The Chromatin Landscape of Colorectal Cancer Cells

Loretta Qinisile Magagula

MGGLOR001

Thesis Presented to THE UNIVERSITY OF CAPE TOWN

February 2020

In fulfilment of the requirements for the degree:

Doctor of Philosophy

Chemical Biology

Faculty : Health Sciences

Department: Integrative Biomedical Sciences

Division: Chemical and Systems Biology

Supervisor: Professor Musa Mhlanga

Co-supervisor: Professor Jane Skok

Declaration

Signed: Signed by candidate

Acknowledgements

It took a village.

Chapter 1 Table of Contents

List of Figures

List of Tables

List of Abbreviations

3C	Chromatin conformation capture
5'-mC	5' methyl cytosine
ALU	Short interspersed repetitive DNA element
aslncRNA	Antisense lncRNA
BAM	Binary Alignment Map
BED	tab-delimited text file
BLAST	Basic Local Alignment Search Too
bp	Base pair
BWA	Burrows-Wheeler Aligner
CAGE	Cap analysis gene expression
Cancer-specific	an observed behaviour is seen only in the context of cancer as compared to wild-type
CBS	CTCF binding site
CCP	Cross-correlation profile
CD4+ T Cells	cluster of differentiation 4 positive T lymphocytes
CGI	CpG island
ChIP	Chromatin immune-precipitation
ChIP-Seq	Chromatin immune-precipitation sequencing
CIMP	CpG island methylator phenotype
CMS	consensus molecular subtype
CNV	Copy number variation
COAD	Colon adenocarcinoma
COSMIC	Catalogue Of Somatic Mutations In Cancer
CpG	CG dinucleotide
CRC	Colorectal Cancer
CRCSC	Colorectal Cancer Subtyping Consortium
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRM	Cis-regulatory module
CTCF	CCCTC-binding factor

DBP	DNA binding protein
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
ENCODE	Encyclopaedias of DNA Elements
eRNA	Enhancer RNA
FANTOM5	functional annotation of the mammalian genome 5
FAP	Familial adenomatous polyposis
FDR	False Discovery Rate
FIMO	Find Individual Motif Occurrence
FISH	Fluorescence In-Situ Hybridization
FRiP	Fraction of reads in peaks
GFF	General Feature Format
GRO-Seq	Global run-on sequencing
GTF	General Transfer Format
GTP	Guanosine Triphosphate
GTPase	Guanosine Triphosphate Hydrolase Enzyme
GWAS	genome-wide association study
HEK293	Human Embryonic Kidney 293
HNPCC	Hereditary nonpolyposis colorectal cancer
IBD	Irritable Bowel Syndrome
IDR	Irreproducible Discovery Rate
IMR90	Human foetal lung myofibroblasts
Kb	kilobase
LAD	Lamina associated domain
LCe	Lower *CTCF* enrichment
LCR	Locus Control Region
LINE-1	Long interspersed nuclear element-1
lncRNA	Long non-coding RNA
LOESS	locally estimated scatterplot smoothing

LOH	Loss of heterozygosity
LS	Lynch Syndrome
MACS	Model-based analysis of ChIP-Seq
Mb	Megabase
MBD	Methyl-CpG-binding domain
MEME	Multiple Em for Motif Elicitation
miRNA	Micro RNA
mRNA	Messenger RNA
MSI	Microsatellite instability
NHGRI-EBI	National Human Genome Research Institute - The European Bioinformatics Institute
NMD	Nonsense mediated decay
NSC	Normalized strand coefficient
ORF	Open reading frame
PA-LCe	Promoter-associated lower *CTCF* enrichment
Pa-lncRNA	Promoter associated lncRNA
PCR	Polymerase chain reaction
piRNA	Piwi-interacting RNA
Pol II	RNA polymerase II
PWM	Position weight matrix
READ	Rectal adenocarcinoma
RIBL	Percentage of reads within blacklists
RIP	Percentage of reads within peaks
RNA	Ribonucleic Acid
RNA-Seq	RNA sequencing
RSC	Relative Strand Correlation
SAM	Sequence Alignment Map
Seq	Sequencing
SICER	Spatial Clustering for Identification of ChIP-Enriched Regions
SOAP	Short Oligonucleotide Alignment Program

SSD	Standard deviation of coverage normalised to total reads
TAD	Topologically associating domain
TALEN	Transcription activator-like effector nucleases
TCGA	The Cancer Genome Atlas
TET	Ten-eleven translocation
TF	Transcription factor
tRNA	Transfer RNA
TSA	Tyramide Signal amplification
UCSC	University of California Santa Crux
UTR	Untranslated region
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Abstract

Chromatin organization is at the heart of deciphering gene regulation as it is instructive to transcription. Current technological advances in next-generation sequencing approaches have offered unprecedented opportunities to interrogate the genomic landscape in multiple pathological and clinical presentations. Historically, mutations and alterations at the genomic loci of protein-coding genes were thought to be exclusively causal to many human diseases. However, the non-coding genome has emerged as the master regulator of chromatin dynamics and transcriptional activity. With cancer increasingly becoming the greatest health epidemic of our time, the comprehensive genomic characterization of tumor genotypes has become central to current therapeutic approaches.

Functioning as the basic unit of chromatin organisation, chromatin loops and topologically associating domains (TADs) compartmentalize genomic loci and their corresponding molecular transcriptional elements in three-dimensional space. Transcription of the human genome is proximity-dependent requiring the cooperative engagement of non-coding elements and epigenetic modifiers to create permissive topological chromatin contacts and structures. The repertoire of chromatin contacts at any given time is regulated by the three-dimensional structure and organization of the chromatin. TAD structures are formed and maintained by chromatin insulating proteins such as CTCF (CCCTC-binding factor) and multi-protein complex, cohesin. The dysfunction of which, through mutational and epigenetic aberrations, directly impacts a plethora of chromatin contacts and the resultant transcriptional profiles within each cell.

Loops and TADs are formed by the binding of CTCF on the conserved 19 bp CTCF binding motif as the chromatin is protruded through the "ring-like" multi-protein complex, cohesin. When two convergently oriented and CTCF enriched CTCF-binding sites (CBSs) come into contact within the ring, cohesin is thought to "hand-cuff" the chromatin resulting in the formation a chromatin loop. These loop structures then serve to compartmentalize and restrict the chromatin contacts and their frequency within each loop. Promoter-resident CBSs can also function as "docking sites" for tissue- and context-specific enhancers.

The dysregulation of CTCF binding has been repeatedly demonstrated to directly alter chromatin contacts in a vast array of cellular contexts including cancer. Fundamentally, CTCF functions as a potent regulator of chromatin contacts, which directly instruct transcriptional status. Thus, CTCF binding has become an attractive regulatory target for manipulating the topological and transcriptional activity of chromatin. In this study, we sought to identify CBSs

with differential, specifically abrogated *CTCF* enrichment that may be hijacked by oncogenes in an attempt to modify transcriptional programmes to favour cancer progression.

To this end, we developed an integrated bioinformatic pipeline to identify promoter-associated lower-CTCF enrichment sites (PA-LCes) in colorectal cancer (CRC) cell lines as compared to primary colonic tissue from *CTCF* ChIP-Seq data. With ever-growing catalogues of next-generation sequencing datasets, including ChIP-Seq, in the public domain, the use of ENCODE datasets proved to be an economical option and added layer of standardization in our analysis. Briefly the pipeline developed in this study takes ENCODE ChIP-Seq FASTQ files from the NCBI SRA using fastqdump as input files. The FASTQ files undergo a quality control and dataset filtration with FASTQC. The filtered datasets are then aligned to the hg38 human genome and fed back into FASTQC to ensure aligned reads pass quality control metrics. The mapped reads are then processed using samtools and duplicate reads are marked with the picard markduplicates argument. Narrow peaks are then called from processed reads using MACS2 and processed using bedtools. Called peaks then undergo a final quality control step using ChIPQCr and are visualized using IGV before undergoing differential enrichment analysis. Differential CTCF enrichment analysis between the peaks in primary sigmoidal colon cells and CRC cell lines is then conducted using DeSeq2 within DiffBind. Lower CTCF enrichment peaks are then used for the discovery of the canonical CTCF MA00139.1 motif using homer and compared to similar annotations in the primary consensus peakset. The resultant lower CTCF enrichment peaks are then annotated using homer and ChiPpeakAnno to determine their genomic locations and extract LCes located proximal (<1kb) to annotated TSS or promoter regions i.e. PA-LCes.

The PA-LCe discovery pipeline developed in this study is highly robust, resulting in some previously validated CBSs implicated in oncogenesis. Intriguingly, the PA-LCe sites identified in this study emanate from bidirectional promoters at oncogenes with differential methylation and transcriptional patterns in cancer. Additionally these PA-LCes transcribe antisense lncRNAs such as the tumor-suppressive aslncRNA ZNF582-AS1. This data adds to the recent body of evidence that suggests that disruption of promoter-associated CBSs leads to fluctuations in promoter activity. Recent studies have implicated the requirement of CTCF-lncRNA complexes at promoter regions in facilitating and regulating CTCF docking on chromatin which subsequently influences transcriptional activity. In accordance with this, our data suggests that the lncRNAs at PA-LCe loci may be molecular targets for the regulation of

CTCF binding and transcriptional activity in CRC. Perturbation of CTCF enrichment at PA-LCes in CRC result in differential chromatin contacts, epigenetic context and, the transcriptional activity of the promoters in which they reside. As CTCF binding at CBSs sites is highly modular, the use of targeted CRISPR-mediated gene-editing and DNA methylation at PA-LCe CBSs may represent viable and druggable oncogenic targets.

Chapter 1 Introduction

1.1. Genome organization

Considered the blueprint of human life, the human genome contains the DNA that controls all the functions within all cell types in our bodies. Although subsets of these cells are morphologically and functionally distinct, the genome in each of the cells is almost identical. Conformational and epigenetic features of the genome are responsible for the differential regulation of the shared genetic information within each tissue and cell. In almost every mammalian cell, the 10 um diameter nucleus houses all the cell's genomic material in non-membrane bound compartments. Mammalian DNA consists of 23 pairs of chromosomes per cell, with each chromosome spanning 59-249 million base pairs cumulating to a length of approximately two meters. The packaging of the cell's genomic material into the compact nuclear volume requires not only a high degree of compaction, but also a system of organization that allows any specific region of the genome to be unpacked, accessed and repacked, efficiently and without knotting, as and when required. This requires sophisticated mechanisms of compaction that also facilitate the accessibility of DNA-relevant machinery for fundamental processes such as DNA replication and transcription. Genome organization and compaction have been evolutionarily conserved in a highly systematic manner in order to retain the functional capacity and interactions within the genome in three-dimensional space.

1.1.1 Chromatin compaction

In its native state the mammalian genome is organized into complex hierarchical high-order structures at multiple scale in three-dimensional nuclear space [1,2]. At the lowest scale, the DNA double helix is tightly coiled around an octamer of core histone proteins to form a nucleosome. Linked by "open" and accessible euchromatic stretches of DNA (compartment B, **Section 1.1.2**), these nucleosomes conform to a structure akin to "beads-on-a-string" (**Figure 1-1**). As compaction ensues, these beads amass to form densely folded chromatin. These fibres are also organized into relatively open euchromatin (compartment A) and condensed heterochromatin (compartment B), based on the post translational modifications of the bound histone. Chromatin fibres further fold into ~1Mb sized sub-chromosomal domains which fold even further to give rise to a single interphase chromosome (**Figure 1-1**). This fractal globule model, first described by Lieberman-Aiden in 2009, allows for the folding and unfolding of chromosomal regions as and when required without the formation of chromatin knots[3].

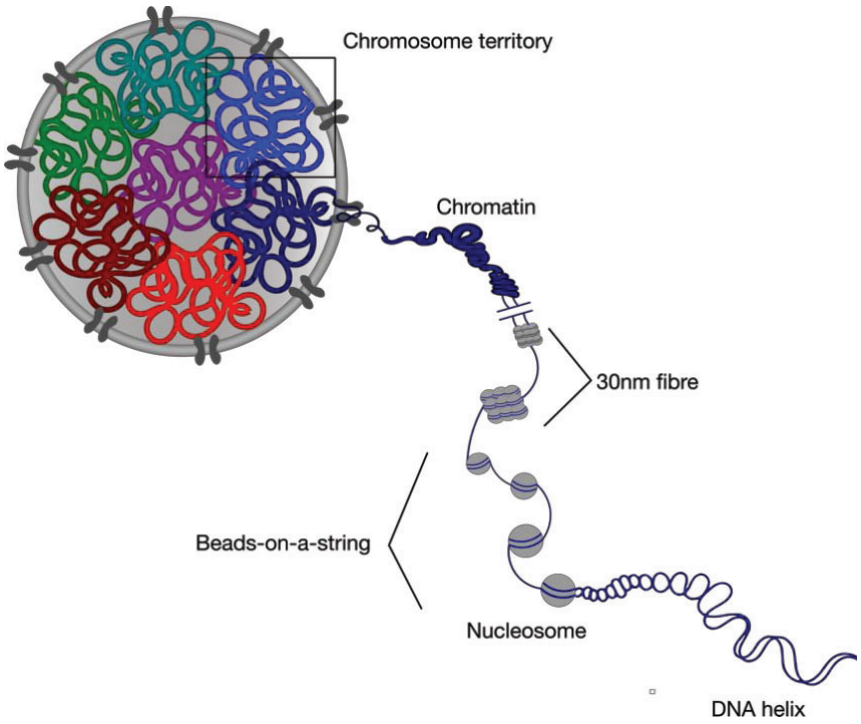


Figure 1-1: Genome compaction within the mammalian nucleus.*The double stranded DNA helix is tightly wrapped around an octamer of histone proteins forming a nucleosome. Nucleosomes and open regions of DNA are arranged in a "beads-on-a-string' structure. Nucleosomes agglomerate to form densely folded chromatin fibres. Chromatin fibres further compact into chromatin that further folds into discrete ~1Mb sized sub-chromosome territories which compact into a single interphase chromosome.*

1.1.2 Chromatin organisation

At the nuclear scale, chromosomes have been shown to occupy distinct spatial territories or domains[1,2] within the nucleus (**Figure 1-1** and **1-2)**. The spatial positioning of these chromosomes has been correlated to chromosome size and gene density, with small gene-dense chromosomes occupying the nuclear interior and larger gene-poor chromosomes localized at the nuclear periphery in a cell type specific manner[4] . Chromosome domains or territories contain megabase-scale compartments of transcriptionally active and inactive chromatin, euchromatin (compartment A) and heterochromatin (compartment B) respectively[2].

Each territory is segmented to form spatially distinct chromatin sub-domains of up to a few megabases in length termed topologically associating domains (TADs), that are enriched with high frequency *cis* or intra-chromosomal contacts (**Figure 1-2**)[3]. TAD structures can be distinguished into phase-separated transcriptionally active and inactive, compartment A and

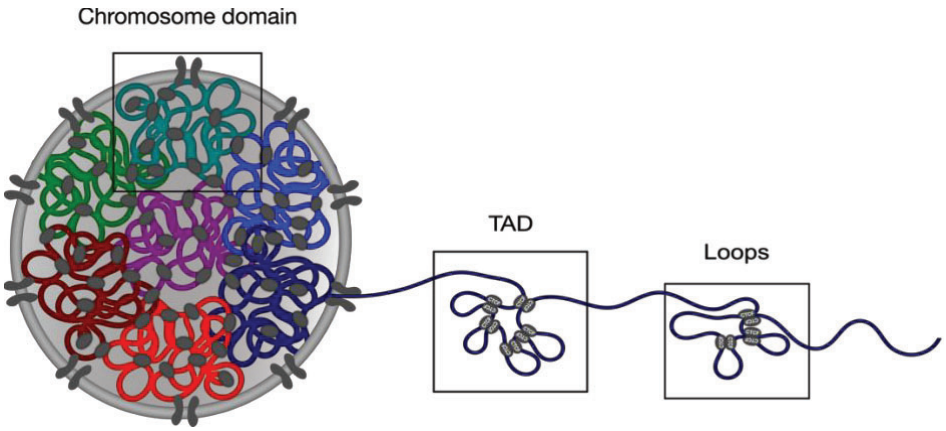


Figure 1-2: ***Chromatin organisation****. Each chromosome in the nucleus is discretely organized into megabase-sized chromosome domains or domains. Within each chromosome domain, chromatin is segmented into a hundred kilobase sized topologically associating domains (TADs) enriched with high frequency chromatin contacts. Each TAD contains chromosome loops that mediate these chroman contacts. The boundaries TAD and loop structures are demarcated, and thought to be formed, by the enrichment of chromatin insulator CTCF.*

B respectively[5]. Heterochromatic or B compartment TADs are typically associated with the nuclear periphery while A compartment TADs are centrally positioned within the nucleus.

Fundamentally, TAD structures serve to constrict functionally related chromatin interactions between enhancers and their target genes (**Section 1.1.2.3.**). These may be up to several megabases apart in linear space, however they are brought together in three dimensional space within TADS thus facilitating transcriptional regulation[5-6]. TADs have been described in many species, suggesting they represent a highly conserved feature of genome organization despite being dynamic structures, particularly throughout development and the cell cycle[6]. These chromatin units define transcriptional regulatory landscapes by compartmentalizing the genome and thus are fundamental in the shaping of functional chromatin organization.

Extensive high resolution Hi-C (chromatin conformation capture) (**Section 1.2**.) maps have demonstrated the presence of intervening sequences demarcating TAD boundaries, which functionally insulate inter-TAD interactions hampering enhancer activity on off-target promoters within neighboring TADs[7] (**Figure 1-2** and **1-12**). A prominent feature at strongly demarcated TAD boundaries is the presence of the conserved CTCF (CCCTC-binding factor) binding motif [7]. CTCF binding sites (CBSs) have been shown to interact with each other in a

highly specific convergent orientation[8]. CTCF is a DNA-binding protein that plays a role in the segmentation and insulation of chromatin[9,10] as well the formation of chromatin loops[11].

Although enriched for CTCF, chromatin loops and TAD boundaries are also populated by the DNA-binding protein complex, cohesin[12]. Cohesin, a multi-protein ring-structured complex, also contributes to the formation of loops and TAD boundary demarcation potentially "handcuffing" CTCF-bound CBS pairs[13]. Currently the mechanisms underlying the formation of CTCF/cohesin-bound TAD boundaries are not fully understood. However, some studies suggest that together CTCF and cohesin are indispensable for the formation and stabilization of TAD boundaries as well as long-range chromatin loops within these domains[14]. Thus, disruptions of chromatin loops and TAD boundaries preventing CTCF and/or cohesin binding, results in aberrant chromatin contact formation and subsequent anomalous transcriptomic activity, often associated with phenotypic abnormalities[15] including oncogenesis[16,17] (**Section 1.6.5**). Chromatin loops share similar features as TADs, with the exception of size (**Figure 1-2**). These are typically short-ranged loops functioning locally, within TADs and spanning less than 2 megabases[9], although some larger loops exist such as the *DXZ4-FIRRE* super-loop[18]. Thus, the current paradigm defines loops and TADs as the basic structural units of multi-scale chromatin organization, that are indispensable orchestrators of complex chromatin regulatory networks and transcriptional activity.

The degree of chromosome intermingling and physical distances between chromosomes within the genome correlates to the level of gene expression. Thus, chromosomes with similar expression levels in are close proximity and/or physical contact in a spatio-temporal manner throughout different cells types and states[19]. These inter-chromosomal contact regions contain transcriptionally co-regulated gene clusters that are enriched with active transcriptional machinery, including phase-separated condensates containing transcription factors and the RNA polymerase II (Pol II) carboxy-terminal domain[20]. Interestingly, the activity of these transcriptional condensates is also governed by a hierarchical order of gene or chromatin docking onto RNA polymerase II[21]. Altogether these observations implicate the four-dimensional chromatin landscape as instructive to transcriptional regulation.

1.1.2 Chromatin regulation

1.1.2.1. Epigenetic markers of chromatin regulation

1.1.2.1.1 Histone modifications

Genome-wide mapping of histone modifications, using Chromatin Immuno-Precipitation Sequencing (ChIP-Seq) (**Section 3.1**) has emerged as a powerful method to interrogate and characterize the functional consequences of chromatin structure. Emanating from hundreds of ChIP-Seq experiments in multiple cell types and disease phenotypes, a "histone code" describing the functionality and repertoire of histone modifications in the regulation of chromatin organisation and transcription has emerged. The "histone code" considerably extends the information potential of the genetic code, where the combination of histone modifications on a particular locus allows for the inference of the type of genomic element, its

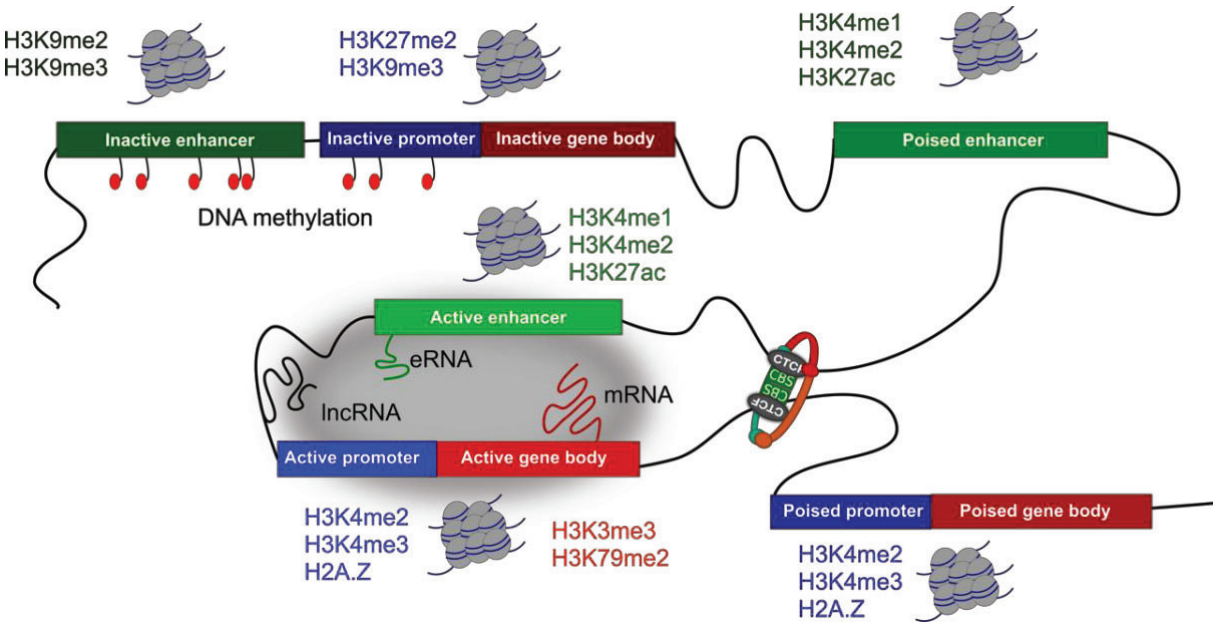


Figure 1-3: ***Histone modifications characterize and demarcate functional elements in the human genome****. Promoters, gene bodies, an enhancer and a boundary element are indicated on a schematic genomic region. Active promoters are commonly marked by histone H3 lysine 4 dimethylation (H3K4me2), lysine 4 trimethylation (H3K4me3) and histone 2A variant (H2A.Z). Transcribed regions are enriched for histone 3 lysine 36 trimethylation (H3K36me3) and lysine 79 dimethylation (H3K79me2). Repressed genes may be located in large domains of histone 3 lysine 9 dimethylation (H3K9me2) and/or trimethylation (H3K9me3) or lysine 27 trimethylation (H3K27me3). Enhancers are relatively enriched for histone 3 lysine 4 monomethylation (H3K4me1), dimethylation (H3K4me2), lysine 27 acetylation (H3K27ac) and the histone acetyltransferase p300. CTCF binds many sites that may function as boundary elements, insulators or structural scaffolds. These features of chromatin help organize and bookmark the DNA for transcriptional activation/repression as well as distinguish functional elements in the large expanse of the genome. RNAPII, RNA polymerase II.*

transcriptional state and its functionality (reviewed in[22]) (**Figure 1-3**). Histone modifications function by regulating chromatin accessibility of transcription factors (TFs), directly mediating the transcriptional state of the marked chromatin and thus, have emerged as potent predictors of transcriptional activity.

Transcriptional state and functionally is pre-determined by epigenetic marks including histone modifications. Determined from thousands of ChIP-Seq experiments over the years, different cell types and models, histone modifications have emerged as a powerful predictor of transcriptional status. Active promoters, enriched for RNA Pol II, are frequently marked by H3K4me2, H3K4me3 and H2A.Z while inactive promoters are enriched for H3K27me3, K3K9me3 and DNA methylation (**Figure 1-3**). Poised promoters, however, are typically bookmarked by H3K4me3, H3K27me3 and H2A.Z (**Figure 1-3**). Similar to promoter regions, gene bodies can be active or inactive, marked by H3K9me2 and H3K9me3 at inactive gene bodies and H3K36me3 and H3K79me2 at active gene bodies (**Figure 1-3**).Active enhancers are frequently enriched with H3K4me1, H3K4me2, H3K27ac and H2A.Z while inactive enhancers are marked by H3K9me2, H3K9me3 and DNA methylation (**Figure 1-3**). Thus, universally H3K4 mono- and dimethylation, H3K27 acetylation and H3K36 trimethylation are markers of active transcription while H3K27 and H3K9 mono- and dimethylation, are markers of transcriptional repression.

1.1.2.1.2 Transcription factors

Transcription factors (TFs) are DBPs that function as “master regulators” of the genomic code. TFs bind specific sequence motifs (**Section 1.1.2.1.4**) on the chromatin to regulate transcriptional activity. Once bound on to chromatin, TFs function as “guides” that promote or hinder the recruitment of chromatin remodelling and/or transcriptional machinery. This is more readily observed in TFs such as PU.1[23] that have the ability to bind inaccessible heterochromatic or nucleosomal DNA establishing regions of “open” chromatin for transcriptional machinery to bind. TFs functioning in this manner have been termed “pioneer factors” due to their unique ability to shape the epigenetic landscape of the cell particularly during differentiation (reviewed in[24]).

The expression and functioning of TFs can be ubiquitous or cell-specific, with many TFs exhibiting lineage-restricted patters of expression, leading to the establishment and maintenance of cell-specific transcriptional programs. Promoter and enhancer sequences contain cell-specific combinations of bound TF binding sites/motifs, which govern their spatio-

temporal transcriptional activity. Thus, the ability of TFs to bind specific DNA sequences alone is considered indicative of their ability to regulate transcription(reviewed in[25])

1.1.2.1.3 DNA methylation

DNA methylation involves the enzymatic addition of a methyl (-CH_3) group onto the 5' carbon of the pyrimidine ring in the cytosine base by DNA methyltransferases (DNMTs). It has been estimated that approximately 3% of cytosines within the human genome are methylated, with almost all 5'-methylcytosines occurring in the context of the CG dinucleotide, CpG[26]. Of these, approximately 80% of CpG dinucleotides in the genome are methylated[27]. Within the genome, unmethylated CpG dinucleotides tend to exist in clusters greater than 200 bp in length, known as CpG islands (CGIs) as well as in regions with large repetitive sequences such as retrotransposons and centromeric repeats[26]. DNA methylation is reversible, primarily through the oxidation of 5'-mC's by Ten-eleven translocation (TET) proteins followed by replication-dependent dilution and/or excision. Furthermore, the absence of DNA methylation maintenance proteins also reverses DNA methylation through dilution during replication[28].

Human genes frequently contain hypomethylated CGIs at their promoters or first exons. The methylation of promotor-associated CGIs inhibits the binding of transcriptional factors and Pol II resulting in the transcriptional repression of said promoter. Furthermore, 5'-mCs can function as docking sites for methylation-dependent DNA binding proteins (DBPs) such as methyl-CpG-binding domain proteins, MBD1, MBD2, MBD3 and Mecp2, which bookmark the chromatin for transcriptional repression by histone deacetylases, polycomb proteins and chromatin remodelling complexes[29]. Aberrant methylation at gene promoters is a ubiquitous feature of oncogenesis. In many cancers, the promoters of tumor suppressor genes tend to be silenced by CGI hypermethylation (**Section 1.3.3**) while CGI hypomethylation tends to occur at repetitive regions of the genome promoting chromosomal instability. Altogether chromatin methylation functions a "bookmark" for transcriptional regulation and simultaneously maintains the structural integrity of the genome.

1.1.2.1.4 Sequence motifs

The regulation of transcriptional machinery on the chromatin is a multi-faceted molecular process governed by; DNA methylation (**Section 1.1.2.1.2.**), histone modifications (**Section 1.1.2.1.1.**), enhancer interactions (**Section 1.1.2.3.**), transcription factor binding (**Section**

1.1.2.1.2), co-factor binding, enhancer RNAs (eRNAs) (**Section 1.1.2.3.2.**) and long non-coding RNAs (lncRNAs) (**Section 1.1.2.2.**). Each of these processes relies, at least in part, on the existence of regulatory sequences or motifs encoded on the bound chromatin. Sequence motifs are broadly classified as short, non-coding, recurring and conserved nucleotide patterns in the DNA sequence for which one or more DBPs, such as TFs, bind with a high affinity. These sequences are typically displayed as sequence logos which represent an underlying "position weight matrix" (PWM), which is the relative preference of the TF for each base at the binding site/motif. Inevitably, variations or mutations, in the nucleotides within a sequence motif directly affects the binding affinity of the DBP or TF on the chromatin, as does mutating the TF itself. This genomic variation has been frequently associated with heritable or acquired disease states including cancer which are catalogued in several large-scale databases including the NHGRI-EBI GWAS catalogue[30].

1.1.2.2. Long non-coding RNAs regulate chromatin organisation and functioning

Simply defined as the transcribed yet untranslated component of the genomic code, non-coding RNAs (ncRNAs) have been classified by base pair length, genomic origin and their functional mechanisms. Representing approximately 70% of the non-coding genome, lncRNAs are described as ncRNA's over 200 nucleotides in length. This size threshold is primarily employed to distinguish lncRNAs from smaller ncRNAs such as micro-RNAs (miRNAs), transfer RNAs (tRNAs) and piwi-interacting RNAs (piRNAs). LncRNAs are classified based on their genomic, epigenomic, tissue and cellular contexts as well as their molecular mechanisms. In the past decade, lncRNAs have been consistently shown to fine-tune transcriptional, translational and post-translational activity. In the context of transcription, lncRNA's typically regulate specific gene targets by altering their epigenetic marks. While the post-transcriptional activities of lncRNAs result in controlling the localization and abundance of their protein targets.

1.1.2.2.1. Genomic classifications of lncRNAs

The genomic classification of lncRNA's is based on the characteristics of their genomic loci (**Figure 1-4**). Long intergenic RNAs (lincRNAs) are transcribed, as the name suggests, from the intervening non-coding portions of the genome between protein coding genes. This class of lncRNAs includes a class of lincRNAs transcribed from enhancer regions (**Section 1.1.2.3**) termed enhancer RNAs (eRNAs). Antisense lncRNA's (aslncRNAs) are transcribed from the opposite strand, and typically in the opposite direction, of a known protein coding gene. These

lncRNAs may or may not overlap, at least in part, with annotated sense genes. Intronic lncRNAs are derived from the introns of protein-coding genes either in the sense or antisense direction (**Figure 1-4**).

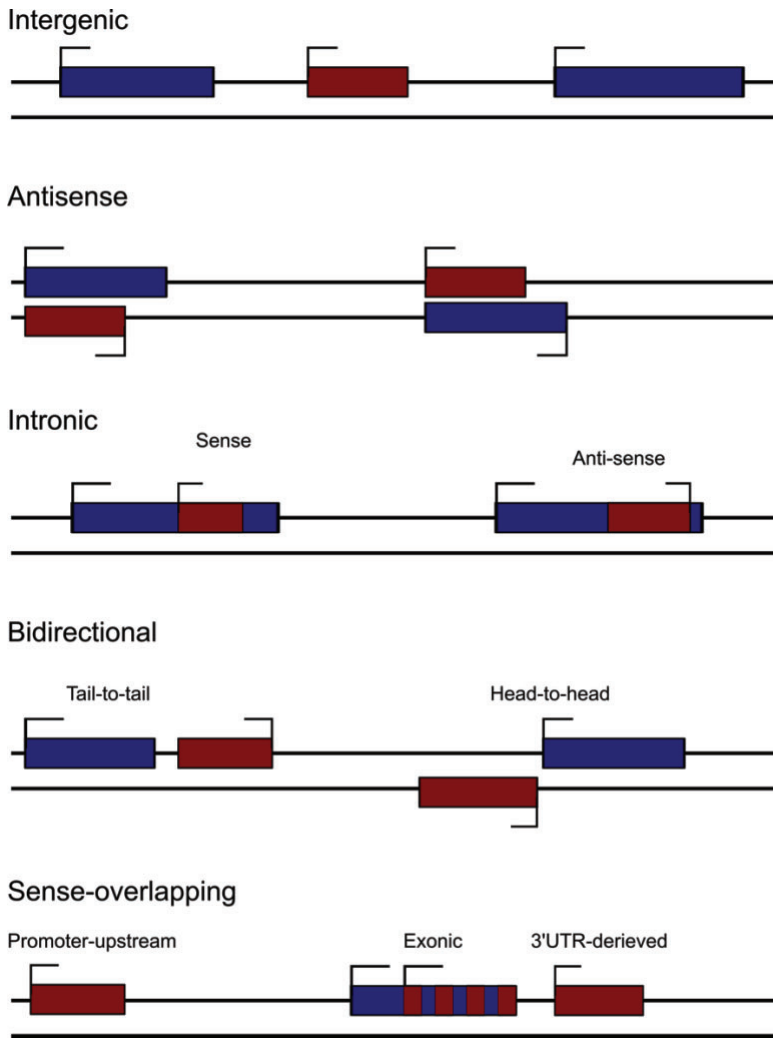


Figure 1-4: ***Genomic characteristics of lncRNAs****. Intergenic lncRNAs are transcribed from non-coding sequences intervening protein coding regions. Antisense lncRNAs are transcribed from the opposite strand of protein coding genes. Intronic lncRNAs emanate from the introns of protein coding genes, these can be sense or antisense. Bidirectional lncRNAs are transcribed either "head-to-head" or "tail-to-tail" approximately 1kb away from protein coding genes. Sense overlapping lncRNAs overlap protein coding regions either upstream from promoter regions, within exons or at 3'UTR loci*[31]*.*

Bidirectional lncRNAs are transcribed "head-to-head" or "tail-to-tail" with protein coding genes within <2 kb. "Head-to-head" lncRNAs transcribed from the same promoter as their proximal protein-coding gene, are also known as promoter-associated lncRNAs (PA-lncRNAs**)**.

Bidirectional lncRNAs include promoter-derived or -associated lncRNAs (pa-lncRNAs) that overlap proximal protein coding promoters either in the sense or antisense direction within 1 kb from protein-coding genes, suggesting these lncRNAs and messenger (mRNAs) are transcribed from the same promoter and potentially by the same transcriptional machinery. Sense-overlapping lncRNAs are considered transcript variants of annotated genes yet are not translated either due to a lack of substantial open reading frames (ORFs), retained introns, or are degraded by nonsense mediated decay (NMD). Within this class there are promoter-derived/upstream lncRNAs, multi-exonic sense lncRNAs as well as 3' untranslated region (UTR)-derived lncRNAs. (**Figure 1-4**). Detailed intricacies of lncRNA definitions are reviewed elsewhere[31–33].

1.1.2.2.2. Molecular mechanisms of lncRNAs

As more lncRNAs are catalogued and functionally characterised, the repertoire of mechanisms by which they regulate their target genes is being refined. Currently, based on their mode of action, lncRNAs are classified as 'decoys', 'sponges', 'scaffolds', 'guides' and

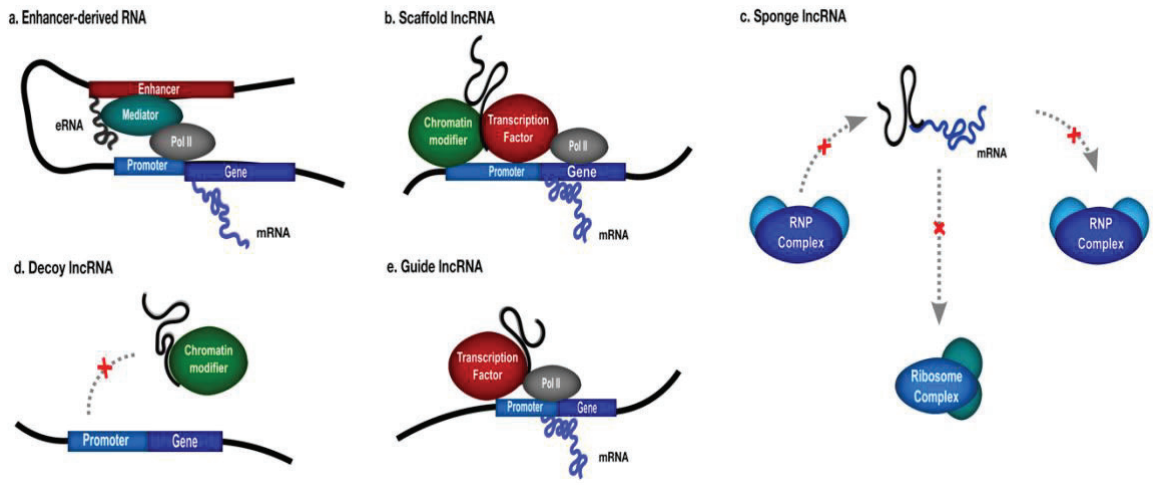


Figure 1-5: Molecular mechanisms of lncRNAs. a. *Enhancer-derived RNAs are short RNAs transcribed from regulatory enhancer regions. eRNAs bind to Mediator, a multiprotein complex, to elicit chromatin looping and long-range chromatin contacts with a parental enhancer and target gene promoters, leading to the activation target genes.* ***b.*** *Scaffold lncRNAs function by binding to chromatin modifiers, transcriptional and/or repressive factors within a single complex, resulting in DNA binding at target gene promoters.* ***c.*** *Sponge lncRNAs function as molecular 'sinks' for miRNAs and/or mRNAs to regulate translation and/ or inhibit the formation of RNP complexes.* ***d.*** *Decoy lncRNAs function as 'sponges' for transcription factors and chromatin modifiers, titrating them away from target gene loci.* ***e.*** *Guide RNAs act as molecular cues for transcription factors and chromatin modifiers at target gene loci[34].*

'enhancer lncRNAs". Decoy lncRNA's regulate gene expression by titrating DBPs such as TFs, away from chromatin. Decoy lncRNAs can also function as molecular 'sponges' or 'sinks' ,binding miRNAs in order to prevent RNAi (RNA inhibition) thus increasing the abundance of the mRNA targeted by the miRNA. Inversely, lncRNA's can 'guide' the proteins they bind to specific DNA or RNA targets.

LncRNAs can also function as molecular 'scaffolds' or 'docks' for one or more DNA-protein and protein-protein interactions. LncRNAs transcribed from enhancer regions (**Section 1.1.2.3**) can function as eRNAs guiding chromosomal looping to exert *cis* gene regulatory effects (**Section 1.1.2.3.2**). eRNAs are transcribed from enhancer regions allowing them to activate genes independent of distance or local genetic context[32,35]. Frequently at the heart of each of the aforementioned molecular mechanisms exhibited by lncRNAs is transcriptional regulation[34] (**Figure 1-5**).

1.1.2.3. Enhancer contacts and activity

Enhancers were first described in 1981 when a tandem 72 bp non-coding genomic sequence, independent of its sequence, orientation or its physical distance from the gene, had the ability to drastically increase the expression of its target gene[36]. This regulatory mechanism has since been a defining feature for transcriptional enhancers. Enhancer functionality is governed by clusters of DNA sequences enriched in TF binding sites, which are occupied by their respective transcriptional activating machinery and histone modifications. Enhancer regions can range from tens of bases to tens of megabases in length, the later appropriately defined as "super enhancers". Enhancers assemble these activating factors into close proximity to their target promoters via enhancer-promoter contacts. This increases the probability and/or the rate of transcription, generally in *cis,* and sometimes over great distances by altering chromatin states and the activity of transcriptional machinery at contacting promoters (reviewed in[37–40]). Thus, enhancers are potent contact-dependent transcriptional activators. However, whether enhancer-promoter contacts are predictive of transcriptional activation and how enhancer-containing loops govern transcription the three dimensional context of nuclear architecture is still enigmatic.

It would appear enhancers make physical contacts with their target promoters to facilitate the docking of transcriptional machinery, though contrary evidence has been demonstrated[41,42]. The absolute requirement for enhancer-promoter contacts in fine-tuning of transcriptional regulation has been highly documented by the Blobel group in the globin locus control region

(LCR) and its promoter targets. The LCR region activates a distinct glass of globin genes in a stage-specific manner throughout erythroid development. By rewiring chromatin contacts, using specific transcription activator-like effector nuclease (TALEN)-mediated tethering of non-target promoters onto the LCR enhancer, the LCR region was shown to function as a potent sequence-independent transcriptional enhancer[43,44]. This locus provides the most compelling and comprehensive evidence to date, supporting the absolute requirement of enhancer-promoter contacts for precise transcriptional regulation. With the advances in chromatin conformation capture (3C)-derived techniques[45] (**Section 1.3.**) including base pair resolution techniques ChIA-Pet[24] and Capture-C[46], several studies have supported this paradigm culminating in the creation of several databases and genome browsers with annotated promoter-enhancer contacts in multiple tissues and linked to several diseases[47–50].

1.1.2.3.1 Epigenetic signatures of enhancer regions

Depending on their epigenetic marks, enhancers can be transcriptionally active, inactive or poised (**Figure 1-3**). Inactive enhancers are associated with repressive marks like H3K27me1, PPRC or DNA methylation (**Figure 1-3**). Poised enhancers, although devoid of nucleosomes, are enriched with the H3K4me1 mark. Active or eRNA-producing enhancers are characterized by H3K4me1, H3K27ac enrichment (**Figure 1-3**)[51]. Active enhancers are highly correlated with the formation of enhancer-promoter contacts. Transcription factor binding, and subsequent histone bookmarks, facilitate docking of transcriptional machinery, along the enhancer locus. As observed in promoters, the docking of transcription factors on enhancers can be cooperative and/or sequential. Enhancers can then “bring along” transcription factors into close proximity to their target promoters thus, acting as potent transcriptional activators.

1.1.2.3.2 Enhancer RNAs

The accumulation of transcriptionally permissive factors, such as TFs, on enhancer regions makes them highly susceptible to transcriptional activation. Thus, it is not surprising that enhancers are pervasively transcribed into typically short ncRNAs, termed enhancer RNAs (eRNAs)[52,53]. eRNAs are currently defined as capped, unspliced, non-polyadenylated, cis-acting non-coding RNAs with a median length of 350 nucleotides. What is intriguing, is that eRNAs shave emerged as potent gene regulators as well. Early studies have linked eRNA activity to their interaction with Mediator, a protein complex that physically “bends” the chromatin to facilitate the formation of enhancer-promoter contacts[35]. Indeed, multiple extragenic transcripts have been found within the classic LCR enhancer. Intriguingly, these

short transcripts, correlating with LCR functionality are expressed in a cell type or differentiation stage specific manner[43]. However, whether enhancer functionality is exclusively governed by the act of transcription, the presence of the eRNA transcripts, or some combination of both, is yet to be determined.

To date, the transcription of eRNAs has been detected in response to various stimuli in multiple phenotypes with the FANTOM5 consortium's latest estimation of human eRNA-producing enhancers currently at 65 000[54]. Notably, eRNAs are typically transcribed at very low levels averaging less than one transcript per cell and as such can only be detected by highly sensitive assays such as 5'-Global run-on (GRO)-Seq Cap Analysis Gene expression (CAGE)-based sequencing and/or Tyramide Signal Amplification (TSA)-Fluorescence In-Situ Hybridization (FISH) microscopy. Attempts to deconvolute eRNA mechanisms have been made, however the results have been varied. A prevailing theme in eRNA functionality, demonstrated by knockdown or knockin experiments, is that eRNAs are potent transcriptional activators that enhance transcription in an expression- and localization- dependent manner. Whether eRNAs, like lncRNAs, can function independently from their parental enhancers remains enigmatic.

With similar epigenetic marks, sequence motifs and transcriptional activity; enhancers and promoters share the ability to integrate spatiotemporal cues in order to co-ordinate tissue-specific gene expression through the expression and binding of tissue specific TFs (**Figure 1-3**). Coupled to that, bidirectional promoters have been shown to function as strong transcriptional enhancers. Bidirectional transcription has become a characteristic feature of active promoters as well as enhancers[55,56]. Recently, a genome-wide characterization of mouse and human promoters revealed that gene promoters, including *FAF2*, *CSDE1* and *TAGLN2*, regulate the expression of several distal genes in an enhancer-like manner[57]. Conversely, intragenic enhancers were shown to act as alternative, tissue-specific lncRNA promoters[57]. Altogether, these discoveries have made the distinctions between enhancers and promoters less apparent

1.1.3 Chromatin insulation

It is evident that chromatin contacts are a fundamental requirement for the transcriptional functionality of the genome. Thus, to prevent aberrant transcriptional activation, the mammalian genome has evolved to form organisational structures, such as TADs and loops (**Section 1.1.2**), in order to locally constrain transcriptional activity in three dimensional space.

The formation and preservation of these structures requires a robust chromatin contact insulation system to prevent aberrant chromatin contacts and anomalous transcriptional activation within the human genome. By insulating the genome into discrete regions, chromatin insulators serve to segregate chromatic regions in the genome in a manner that is

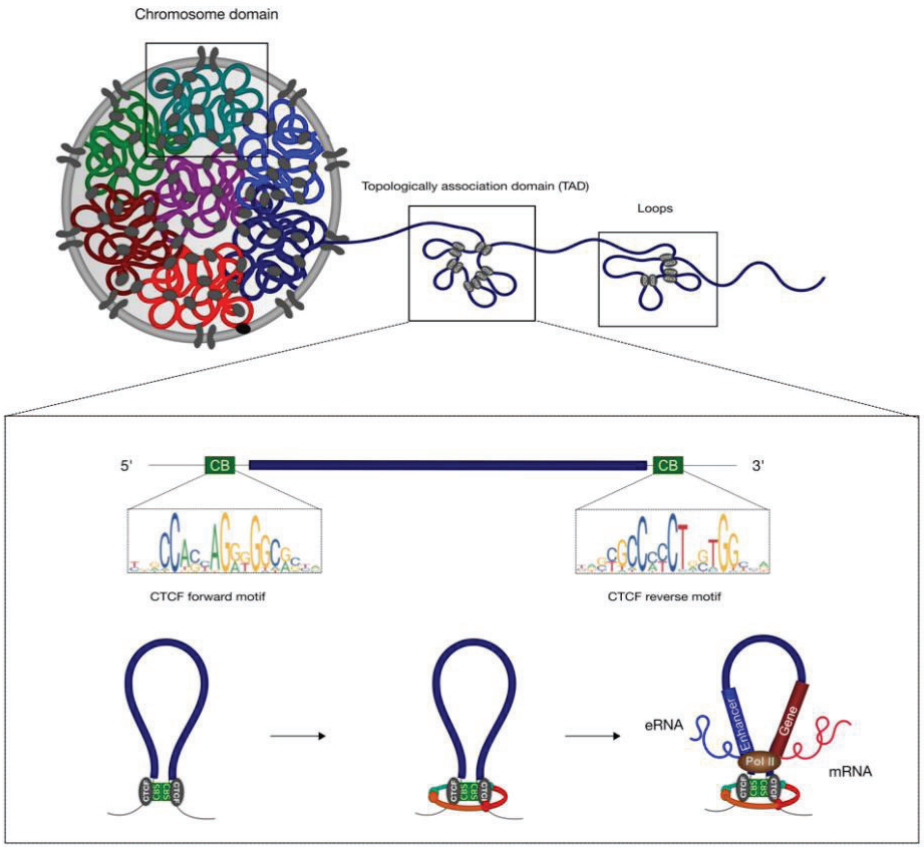


Figure 1-6: ***CTCF loops in transcription. Top panel:*** *mammalian nucleus with chromosome territories (depicted in different colours) are made up of Topologically associating domains (TADs) which encompass several loops. Loop and TAD boarders are enriched with chromatin insulator protein CTCF (grey circles).* ***Bottom panel:*** *The CTCF protein binds to a non-palindromic 19 bp CTCF motif or binding site (CB, green bars), both forward and reverse. Both motifs are displayed here. CTCF binds to CBs as the chromatin is extruded by the "ring-like" multiprotein complex, cohesin (multi-coloured ring). Once two CTCF-bound CBs come into contact, the cohesin ring locks or "hand-cuffs: the CTCF-enriched chromatin contact between the CBs resulting in loop formation. The formation of loops and TADs by CTCF and cohesin allows for high frequency chromatin interactions to occur with the loop or TAD, such as those between enhancer (blue) and promoter (red) regions facilitating loop- or TAD-constricted transcriptional activity insulated from other regions of the chromatin, thus preventing aberrant chromatin contacts and subsequent transcriptional activation.*

permissive to their transcriptional and functional activities. The currently accepted model defining chromatin insulation is known as the loop extrusion model (**Figure 1-6**)[13,58]. In this model, the chromatin is extruded through cohesin, a ring-like complex, until two convergently

orientated and CTCF-enriched CTCF motifs block the cohesin-mediated extrusion. Following which, cohesin is thought to “hand-cuff” this contact resulting in the formation of a chromatin loop (**Figure 1-6**[13,58].

1.2 Characterizing and visualizing chromatin interactions

The dynamic three-dimensional organization of the chromatin is fundamental in regulating and mediating almost all cellular processes. The characterisation and visualization of higher -order chromatin organisation and the epigenome has undergone several technological advances in the last few decades. The techniques used for the detection and characterization of the epigenome are almost exclusively biochemical and are extensively discussed in **Section 3.1**.

1.2.1 Visualizing the spatial organization of the genome using microscopy

Traditionally, genomic conformation was largely based on microscopic techniques such as Fluorescence *In-Situ* Hybridization (FISH), which allows for the evaluation of spatial proximity between genetic loci at diffraction-limited resolution at a single-cell level. This technique can be used to target stretches of DNA or RNA in at single-molecule precision using DNA/RNA probes covalently linked to fluorescent dyes[59,60]. Indeed, this approach was the first to reveal the existence of chromatin territories. The co-localization of FISH probes within a fixed nucleus has been used to determine dynamic, and at times live, genomic loci proximities and transcriptomic activity in single-cells or whole organisms[61,62]. The integration of FISH and immunofluorescence (IF) techniques further revealed the clustering tendencies of active chromatin domains[63]. In the last decade, these techniques have advanced to encompass the automated visualization of the genome at high-throughput, -resolution, -precision in both fixed and live cells.

1.2.2 Chromatin conformation capture techniques

Chromatin looping interactions represent the basic organization structure of chromatin architecture. These chromatin contacts are typically identified using population-based “chromosome conformation capture” (3C) technologies[64]. 3C-derived technologies have revolutionized the field of chromatin organization, enabling the simultaneous detection of genome-wide chromatin contacts. Advances in 3C-derived technologies have led to several paradigm-shifting discoveries. Namely, the compartmentalization of chromosomes into a complex hierarchy of chromatin folding from kilobase range to multi-megabase sized compartments. Developed by Job Dekker and colleagues over ten years ago[64], the 3C technique has evolved into several technologies classified by the types and ranges of chromatin contacts probed.

Briefly, during 3C, a population of cells is chemically fixed with formaldehyde to allow for the formation of covalent bonding between chromatin segments. The cross-linked chromatin is then digested with a 4- or 6-bp restriction enzyme, which defines the resolution of the experiment, where 4-bp cutters yield higher-resolution libraries. The digested chromatin is then ligated to form proximity-based chromatin fragments which undergo reverse cross-linking prior to ligation frequency detection by PCR and/or sequencing[64]

1.3 CTCF binding

1.3.1 The CTCF protein

CTCF is a highly conserved multi-domain protein that binds chromatin in a sequence-specific manner through a tandem array of 11 Cys-His (C2H2) zinc finger domains (ZF)[65]. In classical C2H2 ZF proteins, two cysteine and two histidine amino acids on separate side chains are co-ordinated by a single zinc ion[66]. Typically each ZF, in C2H2 ZF proteins, directly binds a double-stranded tri-nucleotide or "triplet" DNA element. The first high-resolution structure of CTCF ZFs in complex with chromatin, at the human/mouse H19/Igf2 locus, was elucidated by the Cheng lab where the ZFs 3-7 specifically bind to the major groove at the 15-bp core CTCF motif[65] (**Figure1-9**). In the same year, Yin and colleagues resolved the crystal structure of CTCF ZFs in complex with the CBs HS5-1a located on the *PCDHα* enhancer in HEK293T cells[67]. The CTCF ZFs 2, 8 and 9 are highly flexible structures that do not appear to make any base-specific contacts however recent evidence suggests these may play a role in stabilizing CTCF on to chromatin[68]. ZFs 1,10 and 11 do not appear to have a specific function in DNA recognition and thus, are thought to mediate co-factor[65] and RNA[69,70] binding.

CTCF gene and CTCF protein sequences have been shown to be frequently mutated in an array of cancers. Specifically, the *CTCF* gene resides within the minimal deleted 16q22.1 region in several cancers. In colorectal cancer (CRC) specifically, 11.24% of cases had

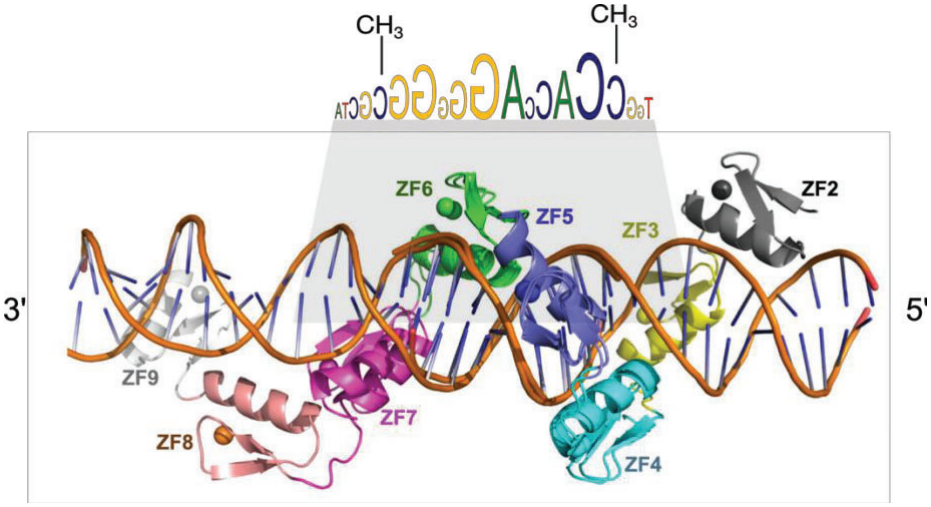


Figure 1-7: CTCF ZF binding on to chromatin[65]. *The zinc fingers 2-9 on the CTCF protein bind to its 15-bp core motif in an opposing orientation, with zinc fingers (ZFs) 3-7 specifically binding to the major grove of DNA, where the 15-bp core motif is located. The flexible ZFs 2,8, 9 do not form base-specific contacts with the chromatin, however the function to stabilize CTCF onto the DNA.*

significant *CTCF* alterations with a mutational frequency of 4.87% in 13 diploid CRC samples reported by TCGA[71]. These mutations span both the ZFs and the N-terminal regions. Specifically, the R339 mutation in ZF3 makes direct contact with C13 towards the 3' end of the core CTCF motif implicating this mutation as a direct effector of CTCF binding onto chromatin[71]. Notably, mutations in other regions within the CTCF protein may, albeit indirectly, affect optimal CTCF binding.

1.3.2 The CTCF motif

The consensus CTCF motif sequence has been a subject of debate for the last thirty years. Identified in the chicken c-*myc* promoter 5'-flanking region, the CTCF motif was first described in 1990, using sequence-specific chromatography[72]. Lobanenkov *et al.* (1990) characterised the CTCF motif as three CCCTC repeats, spaced at 12-13 bp intervals, approximately 200 bp

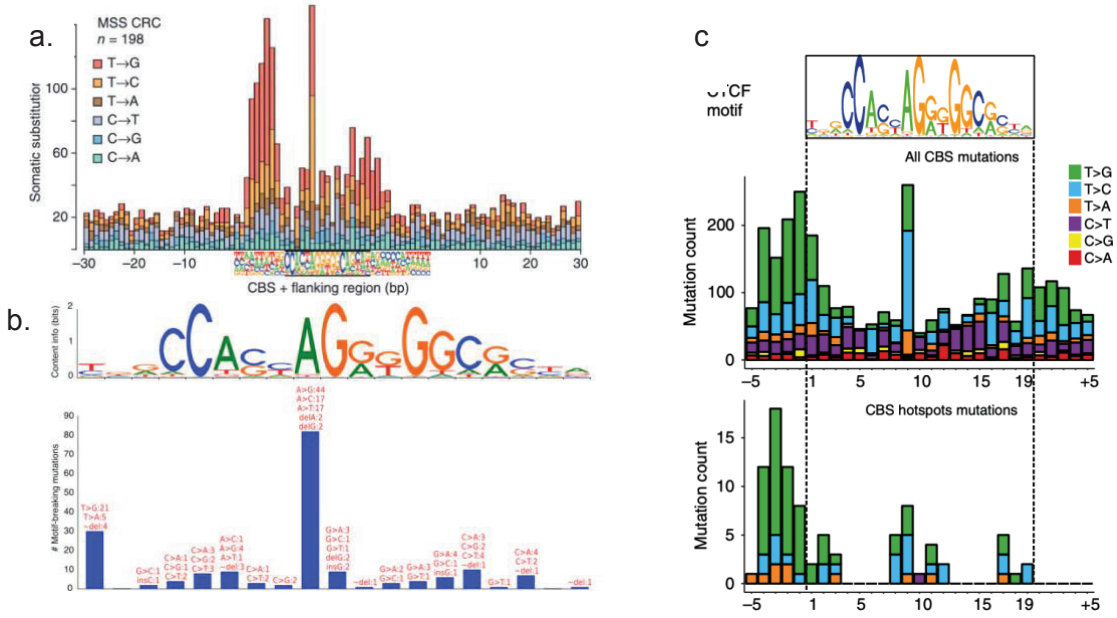


Figure 1-8: Frequent mutations within the CTCF motif in gastric and colorectal cancers[77–79]. a. *The total number of somatic substitutions at the CTCF binding site (CBS) and the 30 bp flanking region, in the 5' and 3' directions in 198 microsatellite-stable CRC Finnish matched tumor samples, as determined by whole genome sequencing mutational analysis by Mutect2 at 28 311 CBSs [78].* ***b.*** *Distribution of the 453 CTCF motif-breaking mutations in 100 Chinese gastric cancer samples[80] analysed by FunSeq2[79]* ***c.*** *Somatic substitution patterns at the CBS and the 5 bp flanking regions, in the 5' and 3' directions, at CBS mutational hotspots and all mutations, respectively. Data was obtained from whole genome sequencing of 40 matched gastric tumor samples from patients from Singapore, 32 ICGC gastric 32 tumors and 100 Chinese gastric cancer samples, called by multiple mutational callers[77].*

upstream from the transcriptional start site (TSS) of the chicken c-*myc* gene. This proposed motif spanned over 45 bp[72]. The deletion of the 110 bp region encompassing the CBS, which also included a Sp1 binding site and a poly(G)-binding protein motif, altered the transcriptional activity of the c-*myc* promoter[72]. In 1993, Klenova and colleagues identified 25 core nucleotides in both DNA strands that were required for CTCF binding at the same promoter[73]. Following the advent of ChIP-Seq technologies, Kim *et al.* identified a core consensus 20 bp motif sequence from over 75% of the 13 804 CBSs in human fibroblasts (IMR90), that displayed high conservation scores with other vertebrates[74]. In 2007, two ChIP-Seq studies identified 13 804[74] and 20 262[75] CBSs, respectively in different cell types namely HeLa, Jurkat and resting CD4+ T cells. The motif discovered by Barski *et al* in 2007, MA0139.1[75] has been validated in multiple cell types, including 56 ENCODE cell lines[76], and is currently accepted as the canonical CTCF motif predictive of CTCF binding (**Figure 1-9**).

As initially described by Lobanenkov, the core CTCF motif (M1) requires additional flanking sequences for tight sequence binding. Over the years, an additional consensus upstream (U or M2) motif has been described. This M2 motif is an 9-10 bp motif, located 5-6 bp upstream from the core motif bound by ZF9-11, which is associated with ~15% of the CBSs in the human genome[81]. Intriguingly, a 6 bp motif downstream from the canonical CBS has also been identified as a destabilizing CTCF binding motif [81].

1.3.2.1 Mutations at the CTCF motif in cancer

Although the CTCF motif has been known to play a fundamental role in directing CTCF binding to the chromatin, the first study to identify CBSs as major somatic mutation hotspots in the non-coding genome was conducted in 2015[78]. In this study, whole genome sequencing (WGS) analysis of 213 primary CRC samples, identified frequent point mutations at ChIP-Seq determined CBSs[78]. It is important to note, enrichment of CBS mutations in this study were described in the context of simultaneous cohesin binding, which are thought to represent only 25% of the CBSs in the human genome. Along with other subsequent studies in gastric, hepatocellular, oesophageal and pancreatic cancers[77–79], CBS mutations in CRC[78] are highly associated with the AT>GC conversion at the adenosine nucleotide in the 9th position (A24) of the CTCF motif[78] (**Figure 1-7**). This A24 base is located at the centre of the fully palindromic core of modules 2-3 of the CTCF motif. The A24 nucleotide is located within at tri-nucleotide bound by Q418 on ZF6, which has been shown to be essential for CTCF binding directionality

(**Figure 1-7**)[67]. CBS mutational studies have identified somatic mutational hotspots in the 5' flanking region, up to 5 bp, of the core CTCF motif in several cancers[77–79] (**Figure 1-8**). These mutations are corelated with genomic instability, chromosomal architecture and transcriptomic aberrations found in cancer[78]. These results highlight the frequency and consequences of CBS mutations in cancer. However, whether these mutations are causative or merely correlated to oncogenesis is yet to be determined.

1.3.3 DNA methylation

It has become increasingly evident that the local genomic context may be as influential on transcriptomic activity as sequence variation. For the CTCF motif, the contextual variation is mediated by pre-existing DNA methylation and the presence of other chromatin binding proteins. Although, global DNA demethylation largely has no effect on global CTCF binding[82], pre-existing DNA methylation at the canonical CBS antagonizes cell-type specific CTCF binding[16,83]. DNA methylation of the canonical CBS sequence has been shown to regulate cell-type specific CTCF binding and its ability to bridge long-distance chromatin looping interactions between distal enhancers and their cognate promoters. The C2 and C12 CpGs on the CTCF motif have been shown to have differential levels of methylation with the later exhibiting higher methylation[84]. Specifically, methylation at C2 of the CTCF motif has been shown to inhibit CTCF binding[78] whereas the methylation of C12 enhances the binding of ZF4-7[65,83].

The cancer-specific hypermethylation at the PDGFRA locus CBSs motifs at abrogates CTCF binding in *IDH* mutant gliomas[16]. Several descriptions of CTCF binding abrogated by DNA methylation have been documented including at the MYC locus[85]. Taken together, this data suggests that CTCF binding, at a subset of cell-specific CBS motifs, is methylation-sensitive and alterations in their methylation status may play a role in oncogenesis. It is important to note that the binding of CTCF onto pre-methylated CpG-poor regions can lead to the initiation and spreading of local demethylation spanning through and beyond the CTCF motif locus[86,87]. Thus far this phenomena has only been characterized at the *PCDHα* promoter where local demethylation is coupled to and driven by the transcription of the *PCDHα* aslncRNA[87].

1.3.4 Enhancer "docking-sites"

Although CTCF does not typically occupy *cis* regulatory elements, when it does occupy these loci, CTCF facilitates enhancer-promoter contacts. An elegant study characterising CTCF

binding on the *MYC* locus by the Young lab established promoter-proximal and CTCF-occupied CBSs as docking sites for multiple cell-specific enhancers[88]. Specifically, the constitutively CTCF-enriched CBS proximal to the *MYC* promoter, in various cancer cell lines, led to the docking of cancer- and cell-specific enhancers at this site. The formation of these enhancer-promoter contacts subsequently leads to the transcriptional activation of the *MYC* gene in a cancer-specific manner. Each of the forementioned contacts were lost upon the abrogation of CTCF occupancy by CRISPR-mediated deletion and/or methylation of the promoter-proximal CBS. Similar results were also observed at the promoters of *TGIF1*, *VEGFA*, *RUNX1* and *PIM1*[88]. This emerging research implicates CTCF in the regulation of enhancer-promoter contacts.

1.3.5 RNA interactions

Recent studies have associated RNA expression and binding to the CTCF protein as a cognate mechanism for establishing a chromatin loops, TADs, promoter-enhancer contacts and intra-TAD transcriptional activity (**Section 1.6.4.2**). CTCF has been shown to bind lncRNAs FIRRE[89] and HOTTIP[90] to promote locus specific CTCF binding while CTCF binding to Jpx has been shown occlude CTCF from chromatin[91]. The *TP53* antisense lncRNA, Wrap53 was also shown to bind directly to CTCF[70], implicating the CTCF-lncRNA binding event in the activation of *TP53* transcription[92]. CTCF-RNA interactions affecting CTCF binding onto chromatin have been demonstrated by alterations in the CTCF-RNA binding domain located at the C-terminus of the CTCF protein[69,70]. Together these findings suggest CTCF-RNA and/or CTCF-lncRNA may play a role in regulating CTCF docking onto chromatin .

Earlier this year, the Maniatis lab demonstrated the requirement of antisense lncRNA transcription in the facilitation of CTCF binding at the *PCDHα* promoter by specifically demethylating said promoter CBSs[87]. This lncRNA-mediated mechanism then allows the HS51 enhancer to bind demethylated protocadherin promoters leading to their transcriptional activation[87]. Furthermore, John Rinn and colleagues demonstrated that the deletion of the Firre lncRNA locus, and not the inter-TAD CTCF boundary site, led to the disruption of the DXZ4 super-loop and its internal interactions in mouse embryonic fibroblasts (MEFs)[89]. Together these studies present a novel CTCF binding mechanism that may be reliant lncRNA transcription, and/or CTCF-RNA interactions to regulate the organisational structure of chromatin and its transcriptional activity. It is important to note that the specificity of CTCF-RNA binding has as of yet not been established.

1.6.5 Loss of inter-TAD CTCF binding alters genome topology

The functional role of inter-TAD CTCF binding on genome topology and transcriptional activation has been a subject of debate with several studies demonstrating somewhat disparate observations. For instance, acute depletion of CTCF leads to a global loss of chromatin loops and TAD structures affecting cell survival with only marginal effects on chromatin contacts and transcription[93,94]. Similar results were obtained upon acute depletion of cohesin in HCT116 cells[95]. Together these studies suggest that alternate mechanisms may be responsible for the establishment and maintenance of transcriptional compartments within TAD structures.

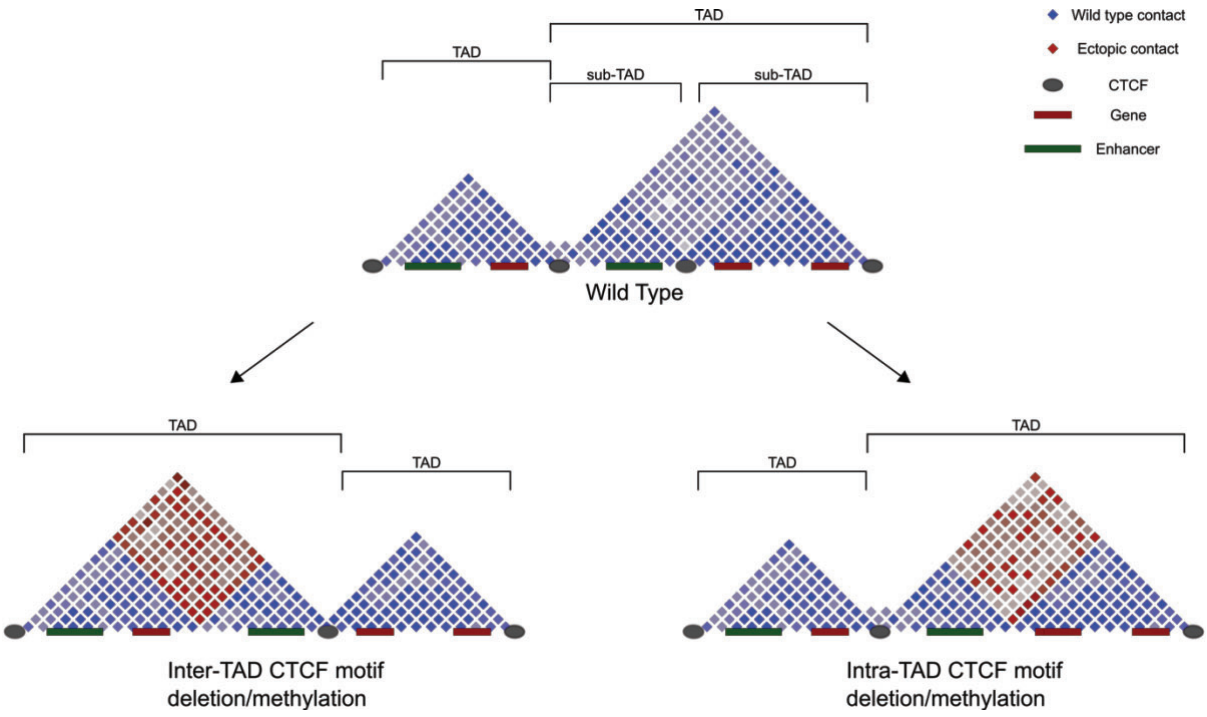


Figure 1-9: Representative Hi-C maps on CTCF motif disruptions affecting CTCF binding including mutations, deletions and/or increased methylation. *An example of a chromatin contact map encompassing two TADs, one of which contains two sub-TADs. (**Top panel**) Inter-TAD CTCF motif disruptions typically lead to the reconfiguration of TAD structures, which results in ectopic chromatin contacts (**Bottom left panel**). Intra-TAD CTCF motif disruptions maintain TAD structure but lead to ectopic intra-TAD chromatin contacts (**Bottom right panel**). Wild type contacts displayed in blue and ectopic contacts in red where color intensity is indicative contact frequency. CTCF sites (grey circles), enhancer regions (green bars) and gene regions (red bars)*

In some studies, deletions or alterations of a single CBS, or a minimal region surrounding a CTCF motif, is sufficient to alter TAD boundaries and rewire promoter-enhancer interactions. A classical study validating this phenomena was conducted by Lupiáñez and colleagues, where CRISPR-mediated structural variations at *Epha4* inter-TAD CBS led to TAD fusions and ectopic contacts between the *Epha4* enhancer, *Pax3* and *Ihh* genes. The loss of these contacts resulted in aberrant expression of these genes as well as phenotypic abnormalities, in mouse embryos[15]. Recently however, a similar study in which serial deletions of inter- and intra-TAD CBSs at the Sox9 and Kcnj2 TADs were formed in the same cell type, revealed that alterations these inter-TAD CBSs alone, did not modify TAD configurations and or transcriptional activity[96]. These apparent discrepancies suggest a redundant system that requires both inter- and intra-TAD CBSs and as well as other additional mechanisms, including cohesin binding and transcriptional compartments, to form internal enhancer-promoter contacts within TADs. Other studies have reported local TAD boundary disruptions occurring only after the deletion of large genomic regions spanning up to 400 kb around discrete inter-TAD CBSs[7,96]. This introduces another conundrum and further re-enforces the lack of exclusivity of CTCF binding at inter-TAD binding sites in mediating chromatin organisation.

1.3.6 Loss of intra-TAD CTCF binding alters promoter-enhancer contacts

Intra-TAD CBSs have been shown to stabilize promoter-enhancer interactions resulting in robust promoter activity and minimal cell-to-cell transcriptional variation[97]. These CTCF-mediated enhancer-promoter interactions, as demonstrated in mouse lymphomas and embryonic stem cells, may be further stabilized by a positive feedback loop through the transcripts arising from participating promoters[97,98]. Intra-TAD CBSs have been shown to function as enhancer “docking sites” in several cancer cell lines leading to cancer-specific transcriptional activity[88]. This was elegantly demonstrated at the MYC oncogene locus, where the CBS located 2 kb upstream of the *MYC* promoter promotes interactions with cancer-specific enhancers in multiple cancer cell lines[88].

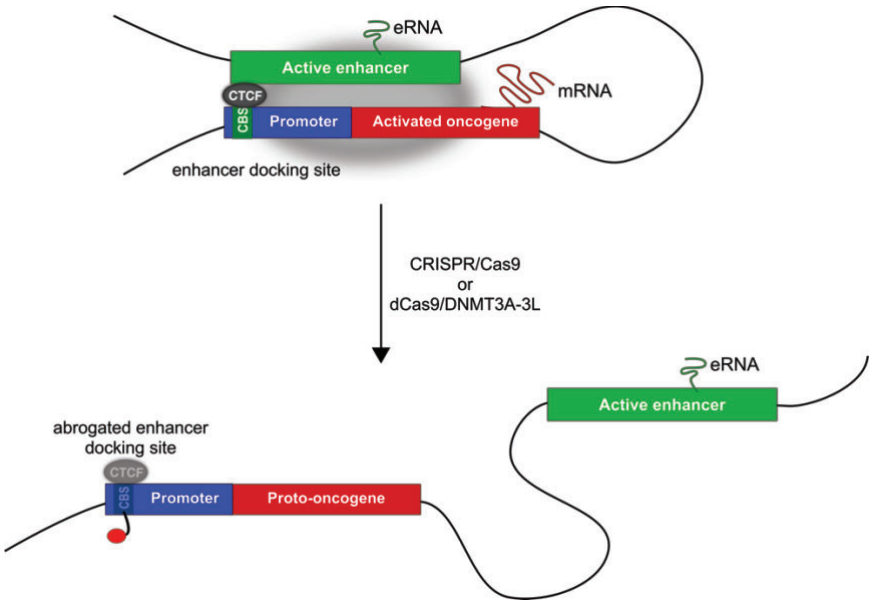


Figure 1-10: Promoter-associated CTCF biding sites function as tissue specific "enhancer" docking sites whose functioning is regulated by CTCF binding. *Abrogated CTCF binding, by methylation (dCas9/DNMT3A-3L) and/or sequence mutation (CRISPR/Cas9) at CTCF enhancer "docking sites" leads to loss of promoter-enhancer contacts leading to a loss of transcription at proto-oncogene promoter region..*

The binding of CTCF at these promoter-associated (PA) intra-TAD CBSs allowed for the "docking" of the *MYC* gene onto cognate cancer-specific enhancer regions resulting in increased MYC promoter activity. Disruption of these PA-intra-TAD CTCF "docking sites" by DNA methylation abrogated enhancer-promoter contacts as well as the transcription of the *MYC* oncogene[88]. Notably, these PA-intra-TAD CBSs are hypomethylated in diverse cancer cell lines, including HCT116, facilitating CTCF binding and transcriptional activity at the *MYC*, *TGIF1*, *VEGFA*, *RUNX1* and *PIM1* promoters[88] **(Figure 1-10)**.

1.2. Epidemiology of colorectal cancer

Globally, fatalities from communicable diseases have decreased significantly however the mortality rates for cancer have increased by over 40% in the past 40 years. In fact, cancer has emerged as the greatest health epidemic of our time. Cancer mortality is currently on the rise and expected to increase by 70% globally in the next ten years[99]. Breast, prostate and colorectal cancer (CRC) are the most prevalent cancers throughout the developed and developing world[100]. CRC predominantly accounts for approximately 10% of cancer-related mortalities in the developed world[101]. CRC is also the third most commonly diagnosed malignancy in men and the second in women, with over 1.8 million new cases reported in 2018[102] and the fourth most leading cause of death globally. In South Africa, where CRC is the fourth most common and sixth most lethal cancer[103], the cumulative risk of CRC incidence rate is currently at 14.4% (**Figure 1-13**)[104].

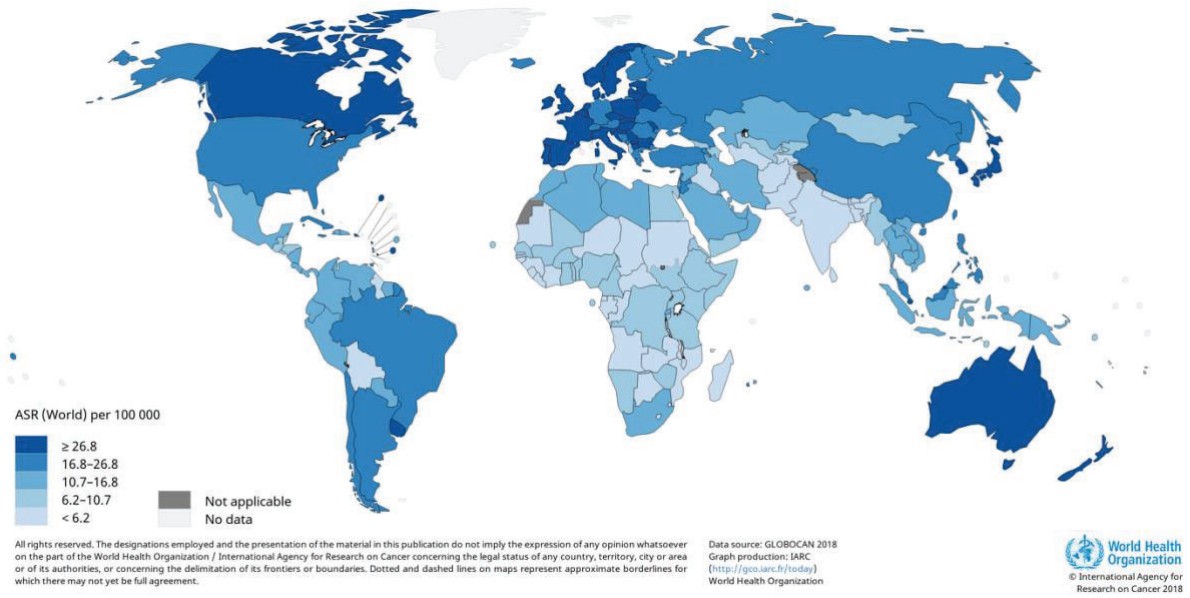


Figure 1-11: Estimated age-standardized incidence rates (World) in 2018, colorectum, both sexes, all ages[104]

Notably, up to 30% of CRC patients have a family history of hereditary colorectal syndromes such as hereditary nonpolyposis colorectal cancer (HNPCC or Lynch syndrome) and familial adenomatous polyposis (FAP)[105]. Most CRC tumors are found in the rectum or sigmoid colon. In the last 50 years, some studies have reported a distal-to-proximal shift (left-to-right shift) in the anatomical distribution of CRC tumors making proximal tumors increasingly common[106].

Microsatelite instability (MSI) and CGI methylator phenotype (CIMP) tumors are more frequent in proximal cancers while chromosomal instability (CIN) cancers are frequently distal[107].

1.3. CRC molecular subtypes

Human cancers are fundamentally heterogeneous with distinct subtypes associated with differences in genetic, molecular, cellular, pathological and clinical proclivities. Coupled to these complexities, are the histopathological and genetic variations observed between tumors arising from the same organ (inter-tumoral) and within individual tumors (intra-tumoral), both across populations and within the same individual. Multiple attempts have been made to investigate, classify and catalogue the diversity observed in CRC tumors for improved diagnosis and therapy. Like most solid tumors, CRC is a heterogeneous disease in which different subtypes can be distinguished by their clinical and/or molecular features. The majority of CRCs are sporadic (70-80%) while approximately 20-30% have a hereditary component due to rare high-risk susceptibility syndromes such as including Lynch Syndrome (LS) (3-4%) and familial adenomatous polyposis (FAP) (~1%)[108]. Notably, a small subset of CRC cases can arise as a consequence of inflammatory bowel diseases (IBD)[109].

Sporadic CRCs develop from one or a combination of the following molecular mechanisms: chromosomal instability (CIN), CpG island methylator phenotype (CIMP) and microsatellite instability (MSI)[102] (**Figure 1-14**). Recently, the International CRC Subtyping Consortium (CRCSC) and The Cancer Genome Atlas (TCGA) introduced a four subtype consensus molecular classification system for CRC tumors (**Figure 1-14**)[110]. Using data from 4151 patients, six CRC subtyping algorithms, integrated datasets and multiple analytical approaches, the consortium described the following consensus molecular subtypes (CMS'):

1. CMS1 MS subtype:
   These tumors are frequently associated with MSI, CIMP and the *BRAF*^{V600E} mutation. Histopathologically, these tumors display a diffuse immune infiltrate and activated JAK-STAT pathways.
2. CMS2 canonical subtype:
   Classified by an epithelial gene-expression signature in the majority of CRC tumors, this subtype it is thought to follow a canonical path of CRC progression (**Section 1.4**) with high CIN levels and activated WNT and SRC signalling.

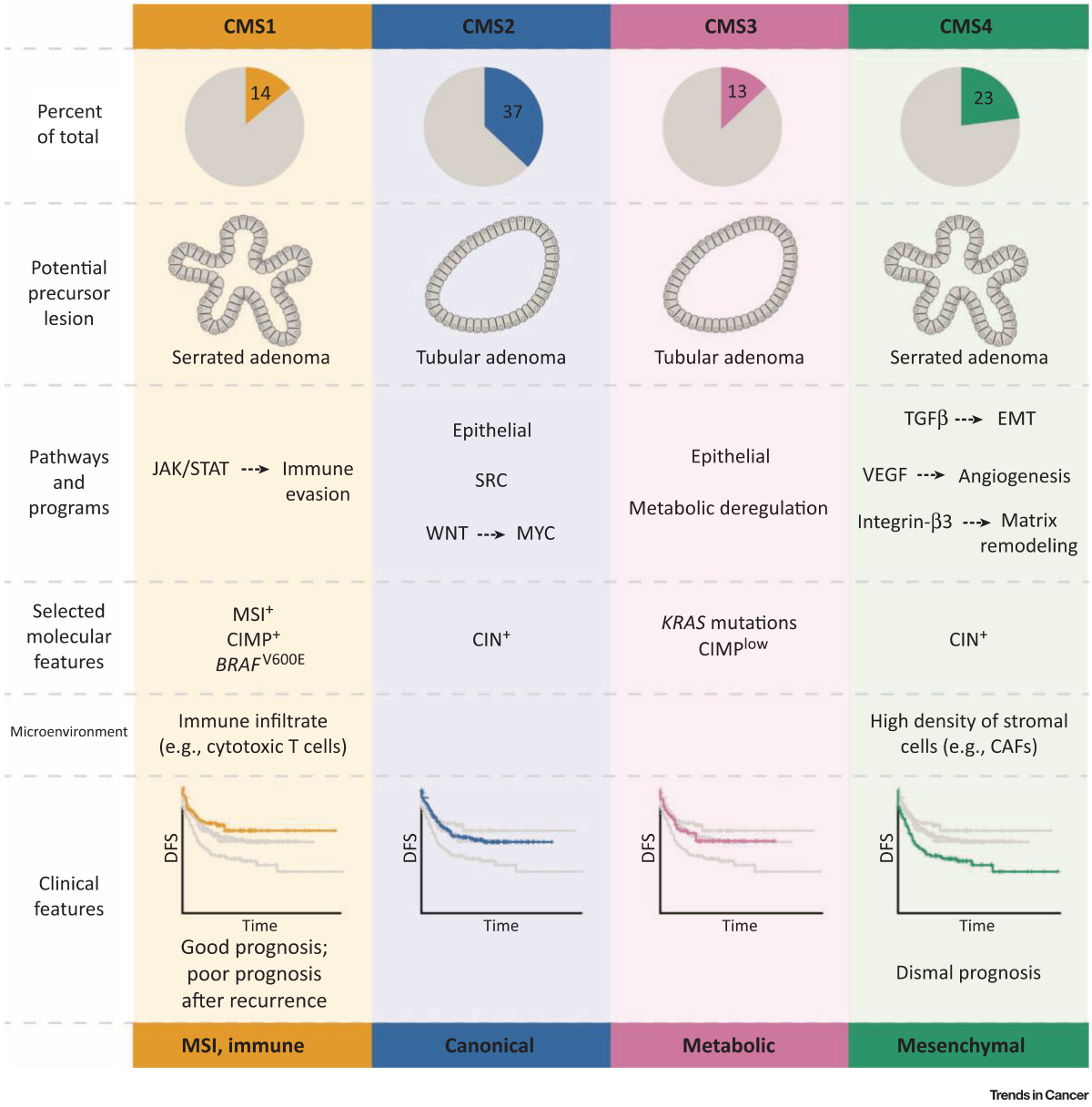


Figure 1-12: Proposed taxonomy of colorectal cancer based on the biological differences observed in gene-expression molecular subtypes*. CIMP; CpG island methylator phenotype. MSI; microsatellite instability. SCNA: Somatic copy number alterations*[110]*.*

3. CMS3 metabolic subtype:

   Based on transcriptomic data, these tumors are characterised by *KRAS* mutations and are associated with deregulated metabolic processes. Like CMS2, CMS3 tumors also display a specific epithelial gene-expression profile with activated integrin-β3, TGF- β and VEGF signalling.

4. CMS4 mesenchymal:

   The upregulated genes in these tumors are integral to signal pathways including: epithelial-mesenchymal transition, angiogenesis, cellular proliferation and extra-cellular matrix remodelling.

Importantly, over 13% of the CRC tumors studied represent either mixed or intermediary subtypes, and as such could not be classified using this system. This suggests more subtypes of CRC are yet to be described. Since the unveiling of the four CRC subtypes, the Catalogue of Somatic Mutations in Cancer (COSMIC) consortium has curated the somatic mutation profiles associated with 7 815 whole exome sequencing (WES) datasets across 26 cancers including CRC. Over 500 colon and rectum adenocarcinoma (COAD and READ) datasets displayed mutations in *BRAF, PIK3CA, KRAS* and *PTEN*, *PIK3CA, TP53 PIK3CA,* and *TP53*[111].

1.4. CRC progression: the adenoma-carcinoma sequence

Cancer is caused by the accumulation of multiple genetic mutations. Typically beginning with a single or combination of oncogene activation which commence the "drive" to cancer. A major hurdle in the identification of these essential "cancer driver" mutations is that most CRC tumors have acquired thousands of MSI or CIN harbouring mutations. In the past, CRC characterization studies have largely focussed on somatic or acquired somatic predictive and prognostic mutational markers in the coding genome. First described by Fearon and Vogelstein over thirty years ago, CRC somatic markers follow the adenoma-carcinoma sequence (**Figure 1-13**)[112].

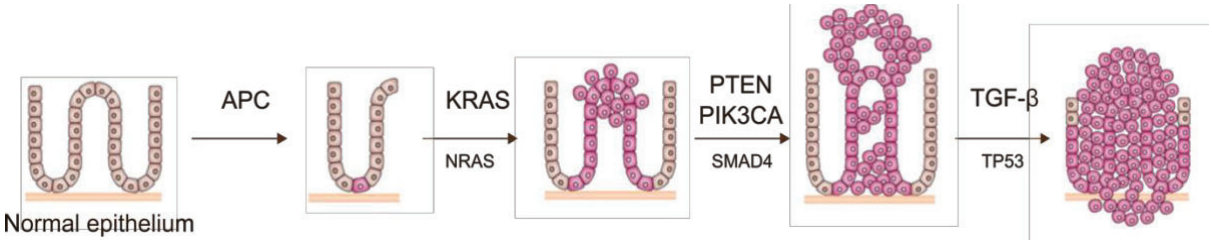


Figure 1-13: Adenoma-carcinoma sequence in intestinal epithelium[101]. *The top panel represents the sequence in classical CIN and MSI cancer. Beginning with mutations in the APC gene, followed by mutations in the KRAS, PTEN/PIK3CA and TGF-β genes in CIMP cancers. Sporadic MSI cancers also begin with mutations in the APC gene and are followed by NRAS, SMAD4 and TP53 mutations.*

The adenoma-carcinoma sequence refers to the progressive and cumulative acquisition of genetic abnormalities that lead to CIN CRC tumors[112]. These include a combination of oncogene activation (e.g. Kristen Rat Sacroma *(KRAS)*, Phosphatidylinositol 3-Kinase catalytic p110-alpha subunit (*PIK3CA*)) and tumor suppressor gene inactivation (e.g. Adenomatous Polyposis Coli (*APC)*, Mothers against decapentaplegic homolog 4 (*SMAD4*) and Tumor Protein 53 (*TP53*)). The mutational aberrations of these genes are now well characterized, and have been shown to involve the dysregulation of five core signalling pathways; WNT/, Ras/mitogen-activated protein kinase (MAPK), tumor protein 53 (p53), phosphoinositide 3-kinase (PI3K), and transforming growth factor beta (TGF-β)/SMAD4 (**Figure 1-13**).

1.4.1. CRC driver genes and pathways

A prevailing paradigm is that CRCs develop by the progressive accumulation of key genetic and epigenetic alterations in colonic-crypt resident cells, that lead to the transformation of normal colonic epithelium to colon adenocarcinomas. CIN sporadic CRCs are associated with gross changes in chromosome number and structure including translocations, gains or losses of chromosomal segments and loss of heterozygosity (LOH). Each of these results in gene copy number variations (CNVs) that affect the expression of genes that regulate cell proliferation, cell cycle check points as well as tumor-associated gene expression. Classical CIN CRC oncogenesis **(Figure 1-13)** is characterized by the sequential accumulation of mutations in the “driver genes”; *APC*[113], *KRAS* and *TP53*[112], which are associated with over 70% of sporadic CRCs . Mutational abnormalities in other loci including; PIK3CA, Deleted in Colorectal Cancer (DCC), SMAD2, SMAD4 and TCFβ have also been reported in CRC, however these mutations only appear to function in the context of pre-existing APC, KRAS and/or TP53 mutations (reviewed in [114]).

1.4.1.1. WNT pathway dysregulation in CRC: APC mutations

Arguably, the earliest event in the adenoma-carcinoma sequence is the mutational inactivation of the *APC* gene, which occurs in over 80% of CRC CSM2 adenomas. The most common mutations in the *APC* gene initiate the formation of benign polyps through the hyperactivation of WNT signalling[115]. APC is a tumor-suppressor and potent inhibitor of the WNT pathway. APC is a component of the Axin-APC degradosome complex that earmarks β-catenin, a WNT effector, for degradation through phosphorylation. DNA hypermethylation at the *APC* promoter has also been implicated with APC inactivation in CRC[116]. The inactivation of *APC* observed in CRC, leads to excess cytoplasmic β-catenin followed by its translocation into the

nucleus[117,118]. Once in the nucleus, β-catenin forms a transcriptional activator complex with TCF4, leading to the transcriptional activation of several genes including the *MYC* proto-oncogene[119]. This leads to a dysregulation of cellular proliferation and differentiation leading to the development of dysplastic crypts (**Figure 1-13**).

1.4.1.2. EGR pathway dysregulation in CRC: KRAS mutations

Proceeding *APC* mutations in crypt-resident epithelial cells are mutations in the *KRAS* gene. Mutations in *KRAS* appear to initiate aberrant crypt foci, an early event in adenoma formation[120] (**Figure 1-13**). KRAS mutations have been found in approximately 50% of CRC cases, the majority of which are associated with the CSM3 molecular signature (**Figure 1-12**). *KRAS* is a proto-oncogene that encodes a small membrane-bound GTPase downstream of the EGFR signalling pathway that activates EGFR ligands including EGF and TGFα[121] Mutations in KRAS, abrogate its GTPase activity, which leads to the constitutive activation of the EGF pathway resulting in incessant cellular proliferation and tumor growth. Activated or GTP-bound KRAS also inhibits apoptosis by activating P13K, which in turn activates AKT, a potent pro-survival kinase that inhibits apoptosis through several mechanisms including inactivating the pro-apoptotic BCL2 proteins[122]. This results in an internal microenvironment conducive for metastasis in the colonic epithelium[123]. Thus, *KRAS* has been established as a potent driver of CRC invasion and metastasis. Unfortunately, KRAS is still considered an undruggable target.

1.4.1.3. TP53 pathway dysregulation in CRC: TP53 mutations

The transition from adenoma to malignant carcinoma is usually associated with mutation and subsequent LOH of TP53 in CRC tumors. *TP53* is a tumor suppressor gene that has been termed as the “guardian of the genome” as it encodes a critical growth suppressor protein that transcriptionally regulates hundreds of genes involved in fundamental cellular processes including DNA repair, senescence, cell cycle arrest, metabolism and apoptosis in response to a number of environmental cues including DNA damage. The mutational dysfunction of *TP53* is considered as a universal hallmark of cancer. In CRC specifically, *TP53* LOH has been reported in 4%–26% of adenomas, 50% of adenomas with invasive foci, and in 50%–75% of CRC tumors[124]. Notably, *TP53* mutational frequencies are higher in distal colon and rectal tumors as compared to proximal rightward tumors[125]. Mutational dysfunctionalities in TP53 result in the loss of its ability to block cellular proliferation, particularly in the context of DNA damage. Thus, loss of TP53 leads to the propagation of damaged DNA into daughter cells

further contributing to mutational selection within the colonic epithelium and promoting the transition from adenomas to carcinomas (**Figure 1-15**)[124].

1.4.1.3. P13K/AKT pathway dysregulation in CRC: PIK3CA mutations

Occurring late in the adenoma-carcinoma sequence, *PIK3CA* mutational dysregulation occurs in approximately 25% of CRC tumors. *PIK3CA* mutations have been frequently associated with poor prognosis in the context of KRAS mutational activation[126,127]. The *PIK3CA* gene encodes for the catalytic p110-alpha subunit of Phosphatidylinositol 3-Kinase (PI3K) alpha, that binds the PI3K regulatory subunits which function in a variety of cellular processes, including cellular proliferation and migration. The PI3K protein family, including PIK3CA, are responsible for the phosphorylation of phosphatidylinositol lipids upon activation by several growth factors including EGF and VEGF. In normal cells, the tumor suppressor PTEN inhibits the PI3K signalling pathway preventing cellular proliferation. However, in numerous tumors where *PTEN* and *PIK3CA* are typically mutationally inactivated, aberrant PI3K responsiveness to EGFR activation results in increased cellular growth and survival. Indeed, EGFR-inhibitors have been shown to have promising therapeutic potential with PI3K inhibition reversing EGFR inhibitor resistance in cancer patients[128]. Thus, together *PTEN/PIK3CA* mutational dysregulation plays a significant role facilitating CRC metastasis (**Figure 1-13**) highlighting it's therapeutic capability.

1.4.2. CRC driver lncRNAs

Most cancer-related mutations occur in regions of the genome outside of genes. Thus, it is unsurprising that oncogenic lncRNA expression, has been implicated in the progression of CRC oncogenesis, similar to the "driver genes" reported. These lncRNAs are summarized in **Table 1.1**.

Table 1-1: LncRNAs implicated in CRC(reviewed in [129–131])

LncRNA	LncRNA classification	Transcriptional expression in CRC	Proposed function in CRC	Potential mechanism
CCAT1-L	Sense	Elevated	Oncogene	Forms long-range interactions with MYC and its enhancers, facilitated by CCAT1-L's binding to *CTCF*
CCAT2		Elevated	Oncogene	Promotes metastasis by being a WNT/ β-catenin effector and activator
MALAT1	Sense	Elevated	Oncogene	Promotes proliferation, invasion and metastasis by regulating WNT/ β-catenin signalling pathway
HOTAIR	Antisense	Elevated	Oncogene	Promotes invasion and metastasis in a PRC2-dependent manner
H19	lincRNA	Elevated or LOI	Oncogene	Promotes proliferation by downregulating H19-derived miR-675, which targets RB.
PVT-1	Sense	Elevated	Oncogene	Promotes antiapoptotic activity through TGF-β signalling
Gas5	Sense	Reduced	Tumor suppressor	Inhibits proliferation and promotes apoptosis by preventing GR binding onto DNA
ncRuPAR	Sense	Reduced	Tumor suppressor	Inhibits tumor growth by downregulating PAR-1
uc.73a	Exonic	Elevated	Oncogene	Promotes proliferation and suppresses apoptosis by an unknown mechanism
CRNDE	Antisense	Elevated	Oncogene	EGFR effector that promotes tumor growth and inhibits apoptosis in a PRC2-dependent manner.
E2F4-AS	Antisense	Elevated	Oncogene	WNT/ β-catenin effector that promotes carcinogenesis by inhibiting cell cycle regulator and tumor suppressive transcription factor, E2F4.
LOC285194/ TUSC7	LincRNA	Reduced	Tumor suppressor	TP53 effector that inhibits tumor growth by downregulating mir211
PTEN-P1	LincRNA	Reduced	Tumor suppressor	Inhibits tumor growth by upregulating tumor suppressive and P13K pathway inhibitor, PTEN through its function as a molecular decoy for PTEN-targeting microRNAs,
CCAL	Unknown	Elevated	Oncogenic	Enhances WNT/ β-catenin signalling by downregulating tumor suppressor AP-2α
MEG3	LincRNA	Reduced	Tumor suppressor	Inhibits proliferation by mediating TP53 signalling
ncNRFR	Sense	Elevated	Oncogene	Promotes malignancy by inhibiting tumor suppressive microRNA let-7
Linc-p21	LincRNA	Reduced	Tumor suppressor	TP53 activator that enhances WNT/ β-catenin signalling
BANCR	LincRNA	Reduced	Tumor suppressor	Inhibits proliferation by targeting cell cycle regulator,CDKN1A and p21

1.5. DNA methylation in the progression of CRC

Globally CRC tumors are hypomethylated compared to normal colonic tissue[132] primarily within repetitive elements such as LINE-1 and ALU elements, which contributes to CRC initiation and increasing genomic instability[133]. However, within this global hypomethylated state, a subset of gene promoters are hypermethylated in CRC, particularly in CIMP tumors. Ordinarily, hypermethylation corresponds with reduced promoter activity due to the lack of transcription factor and subsequently Pol II accessibility at hypermethylated promoters, however hypermethylation can also lead to increased transcriptional activity. In CRC fewer than 10% of methylated genes been shown to have a corresponding decrease in transcriptional activity[134].

Alterations in the methylation status of the core "cancer driver" genes and/or their respective pathways have been reported. The presence and functioning of hypermethylation on the APC promoter is yet to be established. Although, one study has suggested that the APC promoter may be hypermethylated in approximately 20% of CRC tumors[135]. In addition, Several components and targets of the WNT signalling pathway, including *SFRP1* and *LGR5*, have been reported to be hypermethylated (reviewed in[136]). To date, hypermethylation at the *TP53* promoter in CRC tumors has not been studied, however the *TP53* target gene *IGFBP7* has been reported as methylated in CIMP+ CRC tumors, which leads to the inactivation of the *TP53* pathway[137]. Similarly, no effectors or targets of the KRAS pathway have been found to be methylated. Methylation at the *PIK3CA* promoter has been linked to poor prognosis in CRC[138], although the functional mechanism has not yet been established. The P13K/AKT negative regulator, *PTEN* however has been shown to be hypermethylated corresponding to a loss in transcriptional activity in 2% and 20% of MSS and MSI CRC tumors, respectively[139]. In the TGF-β pathway, promoter hypermethylation and downregulation has only been reported in TGF-β regulator *TSP1*[140].

Together these studies suggest DNA methylation is a powerful epigenetic regulator of transcriptional activity. However, the mechanisms of how a specific promoter will respond to DNA methylation i.e. transcriptional repression or activation are currently unclear. DNA methylation at specific gene promoters, particularly at oncogenes, has been repeatedly demonstrated as a powerful prognostic marker. This highlights the need to extensively characterise and distinguish driver and passenger methylation events in CRC tumorigenesis.

Chapter 2 : Study aims and objectives

Aim

Identify and characterise promoter-associated CTCF binding sites with lower CTCF enrichment in CRC

Objectives

1. Develop a ChIP-Seq analysis pipeline, using open-source analysis tools, to identify differential ChIP-Seq enrichment sites.
2. Identifying promoter-associated genomic loci with lower CTCF enrichment (PA-LCe) using ENCODE ChIP-Seq datasets.
3. Identification of canonical CTCF motifs (MA0139.1) at PA-LCe regions in CRC.
4. Characterization of PA-LCe motifs in CRC using genomic annotation data.

Chapter 3 Designing a promoter-associated lower CTCF enrichment (PA-LCe) site discovery pipeline using ChIP-Seq data

It has become increasingly evident that local genomic context influences transcription as strongly as sequence variation. The context of the genomic landscape is mediated by several extragenic factors that bind chromatin and regulate its functional activity. Intragenic factors mediating transcriptional regulation include enhancer regions while extragenic factors can be histone modifications, transcriptional machinery, lncRNAs as well as chromatin-contact insulators such as CTCF. Together these factors mediate transcription primarily through mediating chromatin contacts or "chromatin kisses". The transcriptional status of the interacting sequences can have significant effects on all aspects of cellular functioning, and are therefore tightly controlled. Promiscuous and aberrant chromatin interactions can lead to a host of dysregulated transcriptional activities that can accumulate into systemic phenotypes such as cancer. Thus, detecting and regulating these cancer-driving interactions will advance our ability to diagnose and target these interactions with genomic precision.

Like all physical interactions, a fundamental requirement of chromatin contacts is proximity along the linear genome or through 3D contacts. In mammalian genomes, proximity is regulated by genome segregating proteins and complexes like CTCF, whose primary role is to bind to the chromatin and establish highly specific chromatin interactions that result in the formation of dynamic chromatin loops. Chromatin loops are formed when two convergently oriented and CTCF-occupied motifs interact. This interaction is then "hand-cuffed" by multi-protein complex, cohesin resulting in the formation of a chromatin loop. This loop structure serves to constrain the physical space that the internal loop, or intra-domain, sequences occupy. Chromatin loops also serve to increase the three-dimensional distance between genomic sequences that are located in different loops preventing inter-domain interactions. Chromatin loops can exist within other loops to forming structures known as TADs, whose boundaries are strongly demarcated by CTCF-bound motifs which also serve to constrain chromatin-contacts and regulate transcription (**Section 1.3**).

CTCF binding sites (CBSs) at TAD boundaries, also known as inter-TAD CBSs, have been extensively studied. Early studies suggested that disruptions of inter-TAD CBSs, through the dysregulation of CTCF motif sequences and/or DNA methylation led to the reconfiguration of TAD structures resulting in the formation of ectopic chromatin contacts and alterations in transcriptional activity. However, recent experiments have suggested that the primary effect of disrupting inter-TAD CBSs is the alteration of the chromatin contact landscape. Recently, the focus has shifted to characterizing the role of intra-TAD CBSs in the regulation of contact-dependent transcription. Intriguingly, the disruption of intra-TAD CBSs has been demonstrated to result in the destabilization of promoter-enhancer interactions leading to fluctuations in promoter activity. Intra-TAD CBSs proximal to promoters have emerged as potent transcriptional regulators functioning as "docking sites" (**Section 1.3.4**) for enhancers in a cell- and context-specific manner. It has been shown in some cancers that oncogenes hijack enhancers using this CTCF-mediated docking mechanism resulting in cancer-specific promoter-enhancer contacts and oncogenic transcription[88](**Section 1.3.4**). These promoter-proximal intra-TAD CBSs can be abrogated by genetic and epigenetic editing and thus present intra-TAD CBSs as oncogenic targets. Decreasing CTCF enrichment at these oncogenic promoter-associated intra-CBSs results in the abrogation of cancer-specific contacts and transcriptional activity. Thus, promoter-associated lower CTCF-enrichment (PA-LCe) CBSs could function as diagnostic and therapeutic candidates in cancer.

With CTCF binding playing such a fundamental role in the regulation of chromatin structure and transcriptional activity, we sought to identify differentially enriched promoter-associated CBSs that may play a role in promoting CRC oncogenesis. In this study we define protomer-associated lower CTCF-enrichment CBSs as canonical CTCF motifs within <1kb away from an annotated protomer-TSS with significantly reduced CTCF enrichment in the cancer cell line datasets as compared to the wild type/primary dataset. Only two concrete criterion are required for PA-LCe genomic regions.

1. Lower CTCF enrichment in the cancer dataset as compared to primary/wild-type cells.

2. The PA-LCe genomic region must be <1kb away from the closest annotated TSS in the hg38 human genome.

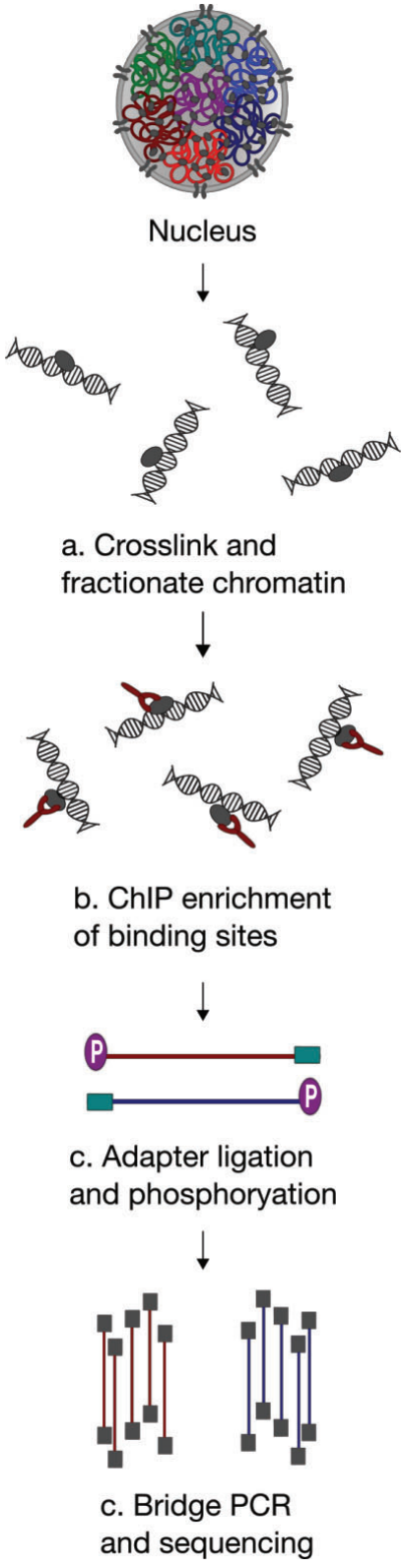
To this end, we developed a PA-LCe CBS discovery pipeline to analyze publicly available CTCF ChIP-Seq datasets in colonic tissues and CRC cancer cell lines. Several challenges with regard to the determination of cancer-specific CBSs exist. The most significant being the availability of primary and cancer patient samples that can be used for analysis as well as the cost of sequencing. However, this is partially mitigated by the increasing repertoire of catalogued and publicly available ChIP-Seq datasets such as those within the ENCODE database. Thus, in this study we employed the use of ENCODE CTCF ChIP-Seq datasets to identify differentially enriched promoter-proximal CBSs between normal colonic tissues and CRC cell lines.

3.1. ChIP-Seq Overview

3.1.1 ChIP-Seq Experiment

Described over three decades ago, chromatin immunoprecipitation (ChIP) has been repeatedly used to directly map protein-DNA interactions *in vivo*[141]. Briefly, in ChIP experiments, an antibody targeting a specific DBP is used to enrich for DNA fragments bound to said DBP (**Figure 3-1**). These enriched sites are then identified and quantified (**Figure 3-1**). ChIP, however, has several limitations most notably the resolution and accuracy in determine CBSs loci, particularly within gigabase-sized vertebrate genomes. Almost twenty years later, these limitations were circumvented by the Wold lab, who performed ChIP with high-throughput DNA-sequencing (ChIP-Seq)[142]. As compared to ChIP, ChIP-enriched DNA fragments in ChIP-Seq are directly sequenced providing higher resolution, low signal-to-noise ratios, greater coverage and fewer artefacts at relatively low cost (**Figure 3-1**). Recently, ChIP-Seq extended techniques such ChIP-exo[143] and ChIP-nexus[144] have been developed to detect DNA-protein binding sites at base-pair resolution, however these techniques have as of yet been applied to a limited number of contexts.

To date, thousands of ChIP-Seq experiments in various contexts have been conducted thus creating a need for recommended experimental standards and guidelines to ensure reproducibility and comparability. The most widely accepted guidelines have been developed by the ENCODE Consortium (**Box 3.1**). In an ideal ChIP-Seq experiment (**Figure 3.1, Box 2.1**), the DNA fragments physically associated with a specific protein are enriched through antibody-based binding. Physical DNA-protein interactions are cross-linked *in vivo* using formaldehyde. The cross-linked chromatin is then sheared by sonication into 200-600 bp fragments. Protein bound and sheared DNA fragments then are enriched by immunoprecipitation, and cross-linking is reversed using high temperature exposure allowing for the release of target DNA, which is then used for the preparation of a sequencing library. The sequencing library is then prepared by a combination of end-repair, phosphorylation and polyA-tailing, index adapter ligation, denaturation and amplification dependant on the sequencing platform used[145] (**Figure 3-1**).



Box 3-1: ENCODE ChIP-Seq Guidelines

I) Standard Measurements for Common ENCODE Cell Types

To ensure consistency in cell cultures, ENCODE has designated common cell types to be used. These include specific culturing instructions, including; cell density, passage number, cell cycle, gene expression, mycoplasma testing and cell freezing standards to be used and recorded for data submission.

II) ENCODE Standards for ChIP-seq Experiments

IIa. Antibody Characterization and Epitope Tagging

To ensure reproducibility of ChIP-Seq data, a set of standards for primary and secondary antibody characterization have been developed. Primary characterization assays include blots, IP Mass spectrometry. Biochemical and bioinformatic secondary characterization assays including 1. RNA interference (RNAi) against target protein, 2. Immunoblotting of epitope-tagged transcription factor, 3. Motif analysis at high quality peaks with an Irreproducible discovery rate (IDR) less than 0.01.

IIb. ChIP-seq Data Production Standards

Sequencing Depth

ChIP-Seq data must be Illumina sequenced with 10-30 million raw reads.

Signal-to-noise ratio

ENCODE endorses a non-redundant fraction > 0.8 for a library of 10 million raw reads.

Figure 3-1: Schematic of general ChIP-Seq experiment[145]. a. *Chromatin is crosslinked and sheared* ***b.*** *Enrichment of DNA binding protein by immunoprecipitation.* ***c.*** *Addition of adapters for end repair and phosphorylation.* ***d.*** *bridge PCR on Illumina ChIP and sequencing.*

3.1.2 ChIP-Seq Analysis

ChIP-Seq analysis standards have also been developed by the ENCODE Consortium (**Box 3-1**) and typically involve the mapping of sequenced reads to a genome assembly, index generation, calling of peaks from mapped reads and the discovery of motifs within called peaks (**Figure 3-2**). Analysis pipelines have evolved from only extracting protein bound regions to differential binding analysis specifically for phenotypically comparable conditions i.e. healthy, diseased and treated conditions. A typical analysis pipeline will follow the ChIP-Seq analysis steps shown in **Figure 3.2**. However, ChIP-Seq analysis workflows and pipelines vary widely in different datasets. To circumvent this variability and the data quality issues that may arise, ENCODE standardized guidelines, practises, quality metrics and tools have been developed to standardize the analysis of replicated and un-replicated ChIP-Seq data (**Box 3.1**).

3.1.2.1 ChIP-Seq Datasets

Despite, the availability of standardized methods, ChIP-Seq analysis is still largely customizable which tends to lead to technical biases. Thus, the choice of downstream ChIP-Seq analysis strategies are dependent on several dataset and sequencing-related factors including, but not limited to:

(1) Sequencing coverage/depth;

Sequencing coverage/depth refers to the average number of unique reads that align to known genome reference bases i.e. number reads X read length/target size. Typically, for a point-source DNA-binding proteins or factors (DBPs), the number of positive ChIP-Seq sites detected increases with the number of sequenced reads. The increased number of reads provides higher statistical power which allows for the ability to detect lower-affinity sites with greater confidence. On the other hand, the number of peaks detected in point-source DBP ChIP-Seq data tends to saturate at approximately 30 million mapped reads, thus current ENCODE standards require that each replicate for point-source DBPs have a minimum of 10 million uniquely mapped reads.

(2) Read length and type,

Typically longer single-end (SE) reads have higher genome alignment rates as compared to short SE reads, largely due to their increased sequence uniqueness[146]. However, for paired-

end (PE) reads, median read lengths appear to improve alignment rates. Notably, PE reads tend to have higher genome coverage rates, particularly over repeat regions, as compared to SE reads of the same length.

(3) Library complexity;

Defined as the fraction of nonredundant DNA fragments in the library (N_{nonred}), library complexity can have significant impact on site discovery and reproducibility in a sequencing experiment[146]. Although, increasing sequencing depth can increase the number of positive sites detected, at some point complexity can be exhausted with the same PCR-amplified fragments repeatedly sequenced to reach greater sequencing depths. This effect can be measured using the PCR bottleneck co-efficient (PBC) which determines whether the distribution of read counts per loci is skewed towards a single read per location. The PBC is defined as the fraction of genomic locations with exactly one unique read versus those covered by at least one unique read[147]. Using this metric, datasets can be classified as having; severe bottlenecking (0≤PBC≥0.5), moderate bottlenecking (0.5≤PBC≥0.8), mild bottlenecking(0≤PBC≥0.5), and no bottlenecking (0≤PBC≥0.5). Samples with severe bottlenecking (0≤PBC≥0.5) values indicate a technical problem with the datasets, such as PCR bias.

ENCODE uses the Non-Redundant Fraction mapped reads (NRF) scores, $0.5 \leq \mathrm{NRF} < 0.9$, as a library complexity quality control metric for point source DBPs. Simply put, the NRF is defined as the ratio between the number of distinct uniquely mapped reads and the total number of reads. It is important to note that increasing sequencing depths results in lower NRF scores.

(4) The availability of a control or input sample.

The non-uniform DNA breakage during sonication in ChIP-Seq experiments, caused by the overrepresentation of open regions of chromatin underpins the need for a control or input sample for ChIP-Seq analysis. Furthermore, aneuploid cell lines with large genomic duplications can influence called peak sizes and rankings. The use of control or input samples. therefore substantially alleviates these biases. Control samples typically involve the use of a non-nuclear antigen binding control antibody such as Immunoglobulin (IgG) whereas input samples do not undergo immunoprecipitation and all DNA fragments sequenced.

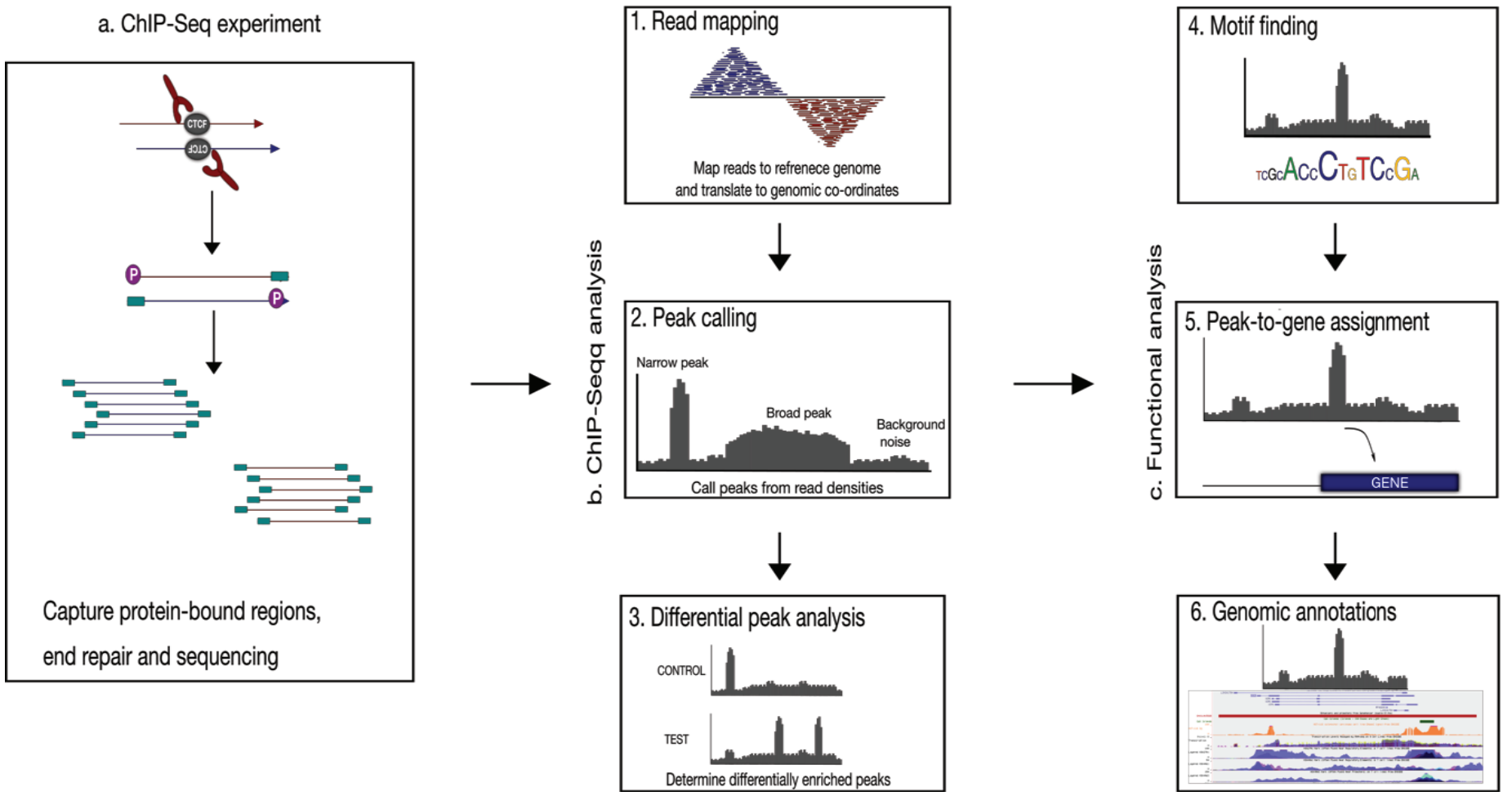


Figure 3-2: A typical ChIP-Seq Analysis Pipelinea. *The workflow of a ChIP-Seq experiment which involves the immunoprecipitation of protein-bound DNA using a corresponding antibody, followed by end repair and sequencing.* ***b****. ChIP-Seq analysis typically involves three main steps: 1. Mapping sequenced reads to a reference genome; 2. Calling peaks from read densities; 3. Differential peak calling to determine differentially enriched peaks between two or more conditions.* ***c****. Functional analysis of called peaks can be conducted be conducted in silico by 4. De novo motif finding; 5. Peak-to-gene assignments based on linear motif proximities to annotated genes; 6. Genomic annotations of CTCF motif loci using epigenetic signatures*

(5) The availability of replicates

As with most experimental approaches, the use of at least two replicates ensures greater reproducibility and data confidence. Although ENCODE recommends at least two biological replicates for ChIP-Sequencing in cases where the availability of experimental material, such as patient biopsy samples, can be exempted however in such cases the data has limited value and data confidence.

(6) The quality of the antibody used

Arguably, the most influential technical feature on the quality of the ChIP-Seq data produced is the antibody to capture the chromatin during the first step the ChIP-Seq experiment. The affinity, specificity and cross-reactivity of the antibody used in a ChIP-Seq experiment can introduce significant technical bias thus affecting sequencing depth. The same experiment using the different antibodies targeted to the same protein, or even using the same antibody in different samples can produce completely different peak distributions. As a result, the ENCODE Consortium has implemented multiple TF antibody characterization methods and thresholds in an attempt to mitigate the these technical bias. However, it is important to note that despite each of these experimental validations/checkpoints, biases caused by the antibody are inherent to the ChIP-Seq technique and cannot be completely circumvented.

3.1.2.2 ChIP-Seq Analysis tools

3.1.2.2.1 Read quality metrics

Prior to mapping sequenced reads to the reference genome of choice, reads are typically filtered by applying a several quality cut offs. A report detailing the raw read quality, sequencing errors or biases can be obtained by running quality control algorithms such as FastQC[148]. FastQC assess GC content, the over-abundance of adaptors and over-represented sequences which may indicate PCR duplication rates. Phred quality scores, denoted as MAPQ, can also be used to describe base call confidences in each sequence tag. These scores infer error probabilities and can be used to inform the filtration of low-quality reads and read trimming. A base with a MAPQ of 50 for example, means that there is only a 1 in 100 000 chance that base has been called incorrectly. A widely accepted Phred base score is typically ≥30.

3.1.2.2.1 Read mapping

Arguably, the most crucial early stage of sequencing data analysis is DNA read alignment/mapping. Briefly, read mapping refers to the process of aligning or mapping sequenced reads (FASTA or FASTQ) onto a reference genome in order to determine the genomic loci from which read was sequenced.

Currently, the most popular read alignment tools for mammalian genomes are Bowtie2 [149] and BWA[150]. Recent aligners such as Bowtie2, support gapped, local and pair-end alignments. Such aligners allow a settable number of mismatches in the reads. The preferred percentage for uniquely mapped reads reported by an aligner for ChIP-Seq data, although variable dependant on the organism, is typically above 70%[147]. Percentages lower than this, may suggest excessive PCR amplification and inadequate read or sequencing errors. High numbers of non-uniquely mapped reads may also suggest the ChIP'ed protein binds to repetitive DNA sequences, this can be circumvented by using PE sequencing or longer reads to reduce mapping ambiguity. However, most peak-calling algorithms tend to filter out multi-mapping reads during binding site discovery.

Table 3-1: Short read alignment tools

Software Tool	Seeding strategy	Seeding matching	Source	Extension	Notes
BWA[150]	MEM	Exact	Burrows-Wheeler transform	BLAST-like	Fast, efficient, based on Burrows-Wheeler transform
Bowtie[151]	k-mer	Inexact	Burrows-Wheeler transform	Global and local	Similar to BWA, part of suite of tools that includes TopHat and CuffLinks for RNA-Seq processing
GSNAP[152] (SNAP)	q-gram (k-mer based)	Inexact	Hash table or q-gram index	Global	Considers set of input variant alleles to better align to heterozygous sites
SOAP2[153]	k-mer	Inexact	FM-index	BLAST-like	k-mer inexact match seed; support at most 3 mismatches; GPU calculation supported
Novoalign[154]	k-mer	Inexact	Hash table	Global	Similar to GSNAP and supports mismatches and gaps of up to 50% of read length

3.1.2.2.2 Alignment file processing: SAMtools

The above-mentioned alignment tools typically generate genome alignments in various formats which can complicate downstream processing. Thus, in 2009 the Wellcome Trust Sanger Institute designed the Sequence Alignment/MAP (SAM) format which supports all sequence types an alignments, and the corresponding software package SAMtools[155]. The mandatory features in the SAM format are shown in (**Table 5.3)**. SAMtools is a genomic library and software package that is used to manipulate and parse genomic alignments in the SAM/BAM formats (**Tables 5.3-5.4**). Features in SAMtools include, but are not limited to alignment format conversion, sorting and merging alignments, removing PCR duplicates and generating base pair position information, calling SNPs as well as short indel variants.

3.1.2.2.2 Peak calling

An essential step in ChIP-Seq analysis is the detection of peaks from ChIPed regions, above background or input samples. Peak calling algorithms are used to identify regions of ChIP enrichment i.e. significant number of mapped reads at genomic region. This is arguably, the most pivotal analysis step for ChIP-Seq data. Several peak calling algorithms and software

packages have been developed; including MACS[156], Phantompeakqualtools[157], HOMER[158], SPP[157] (**Table 3-4**). Peak callers differ in their signal smoothing, background modelling and normalization methods. For point-source DBPs, peak callers such as SPP and MACS2 use cross-correlation shape based methods to model the strand-specific spatial distribution of reads and peaks are called when the distribution of mapped reads conforms to a probabilistic distribution around the binding site.

3.1.2.2.2.1 Identifying ChIP-Seq enrichment

ChIP-Seq binding site detection can be performed using one of three approaches; (1) Peak-finding, (2) Peak-pairing or (3) Probabilistic binding detection **(Figure 3-4)**[159].

3.1.2.2.2.1.1. Peak-finding

ChIP-Seq peak-finding analysis methods such as Model-based Analysis of ChIP-Seq (MACS)[156] and SICER[160] estimate a fragment size (d), which is then used extend reads to represent the original ChIPed fragments. Approximately 1000 randomly sampled regions, with 10- to 30-fold enrichment above genome background, are then used a model peaks.

Using this peak model, MACS then slides a window size of $2d$ across the genome to identify regions that are significantly enriched relative to the genome background. Overlapping windows are then merged to form candidate peak regions. The Poisson distribution with dynamic parameter λ_{local} for candidate peaks is then calculated and varies along the genome i.e. the λ_{local} value for a specific locus is defined as:

$$\lambda_{local} = max(\lambda_{BG}, [\lambda_{region}, \lambda_{1k}], \lambda_{5k}, \lambda_{10k})$$

where, λ_{BG}= constant estimated from genome background

λ_{region}= constant estimated from the candidate region

λ_x= constant estimated from an x-bp window centred at candidate region in control sample

In most cases, the ChIP-Seq and control samples have different sequencing depths thus, MACS uses the relative difference (*de novo* normalizationto scale reads in both samples. In cases where control samples are unavailable, λ_{local} is calculate from the ChIP-Seq sample excluding the λ_{region} and λ_{1k}. MACS then uses the calculated λ_{local} value to assign an enrich-

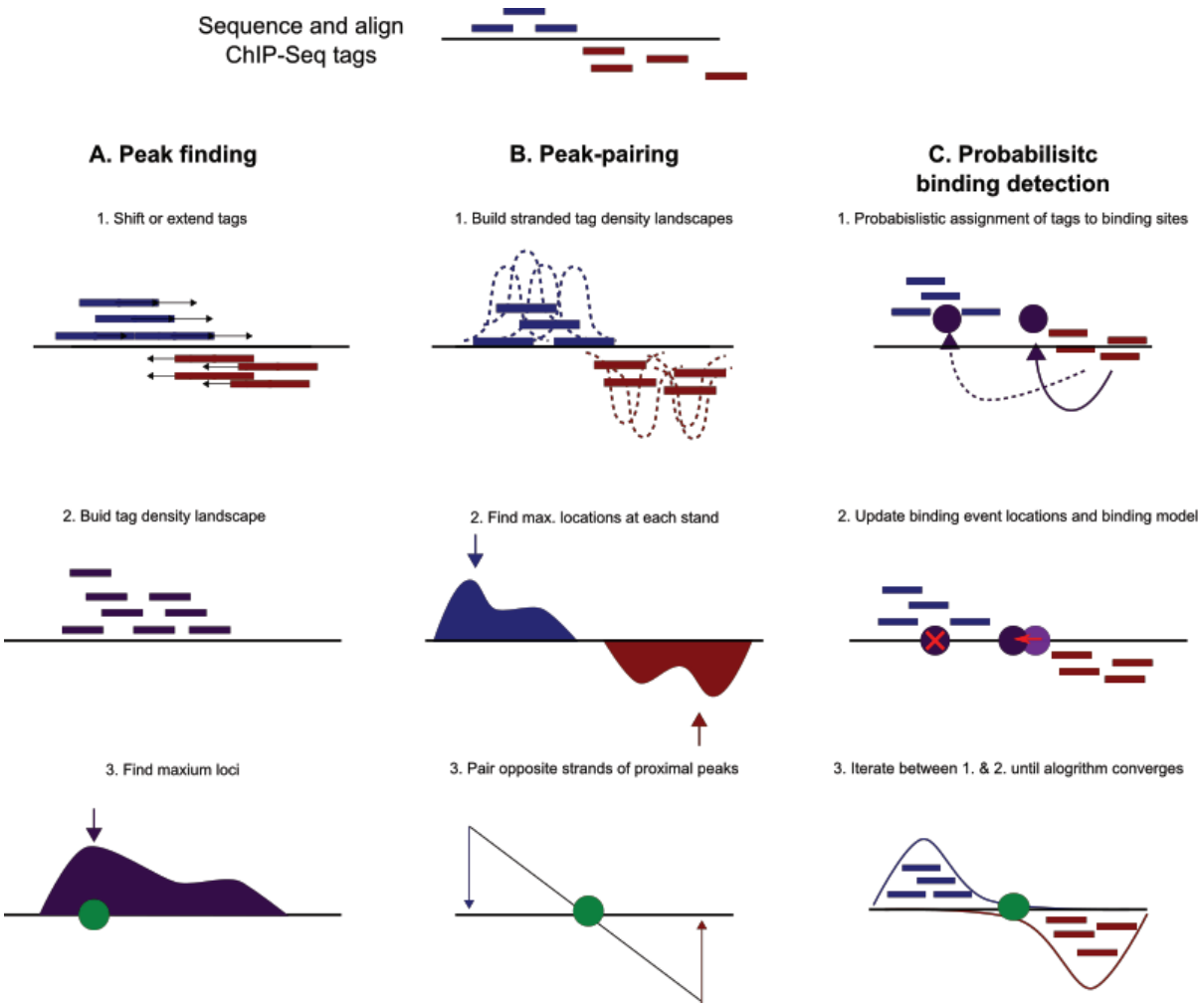


Figure 3-3: Peak finding approaches[159]***. a****. Peak finding methods either* ***1.*** *shift or extend ChIP-Seq tags locations in a 3' direction by half the expected fragment length or equal to the expect fragment length, respectively.* ***2.*** *Opposite tags are then merged to build unstranded tag density landscape which allows for the prediction of binding site locations based on the maximum tag density location.* ***b.*** *Peak-pairing methods* ***1.*** *build similar, albeit stranded, tag density landscapes that neither shift or extend.* ***2.*** *Peak maxima are predicted separately and* ***3.*** *proximal peaks on opposite strands are paired and binding events are predicted from peak-pair mid-points.* ***c.*** *Probabilistic binding detection methods begin by* ***1.*** *training by initial binding event location guesses, which are model how tags are expected to be distributed around "real" ChIP-Seq binding events. For every training step, each ChIP-Seq tag is probabilistically associated with proximal binding events that are* ***2.*** *updated to fit associated tags and the binding model is updated to accommodate the accumulation of tags around all current binding events.* ***3.*** *Steps 1 and 2 are iterated and binding events with a low number of assigned tags are filtered out until the model converges to a final set of binding events. Purple circles represent predicted binding sites during peak finding. Green circles represent final and program output binding sites.* [159]

-ment *p*-value to candidate peaks, which are then filtered by a *p*-value ≤ 10^{-5} threshold to determine final peaks. In the event that a control sample is available, for a particular *p*-value threshold, MACS calculates and thresholds peaks using the empirical FDR.

3.1.2.2.2.2.Peak-pairing

Peak-pairing analysis methods such as those used in GeneTrack[161] and Model-based Analysis of ChIP-exo (MACE)[162], build similar tag density landscapes to MACS, without shifting or extending tag locations. These algorithms determine peak locations on each strand separately. Nearby peaks in appropriate stranded orientations and within a given distance are then paired. While MACE functions by creating peak-pairs of closely spaced peaks on opposite strands, GeneTrack requires peaks to be paired manually. These algorithms then iterate the probabilistic assignments and update the binding event locations and model until the algorithm converges to a final set of binding locations, which are predicted from the peak-pair midpoint (**Figure 3-4**).

3.1.2.2.2.3.Probablistic binding detection

Peak detection analysis methods such as Genome Position Systems (GPS)[163] and PeakSeq[164] build probabilistic binding models. Initially, GPS summarizes the observed spatial read distributions from the control or input ChIP-Seq samples by assuming that every binding event will produce the same distribution of reads. This approach can only be used when an input or control sample is available. Next, GPS uses a probabilistic mixture model to assign binding event probability to the genome at single-base pair resolution. The number of reads assigned to a base by the mixture model is then used as a measure of relative strength of a predicted event at the base in question. Finally, GPS filters discovered events by comparing the number of reads at the predicted events to the corresponding normalized number of reads in the control/input sample (**Figure 3-4**). The statistical significance of the discovered events is then computed using a binomial distribution[164], which is corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure[165].

The peak calling algorithms (**Table 3-4**), use various statistical approaches to call and rank ChIPed peaks when comparing control/input and ChIPed samples. An intuitive and widely-used linear normalization technique between ChIPed and control samples is based on sequencing depth. Using this scaling method, the number of reads within each sample is multiplied by a scale to make the total number of reads across samples the same or comparable. Typically, the output of such algorithms produces and rank regions of enrichment based on absolute signal or computed significance of enrichment, P-values, q-values, fold enrichment and/or FDRs, which can be influenced by the statistical model used, sequencing depth or the abundance of binding sites in the genome.

To assess for the reproducibility of reads and called peaks, the Pearson correlation coefficient of mapped read counts at each genomic position can be computed for each replicate. Replicate samples typically have a Pearson correlation >0.9 and values lower than 0.5 suggest unrelated samples or one of the replicates may be of low quality. An informative measure of reproducibility can also be conducted using is the irreproducible discovery rate (IDR)[166] which indicates the signal-to-noise ratio and assigns each signal a reproducibility index for each peak. IDR values >0.05 indicate consistent called peaks between replicates.

Table 3-2: Features of common peak-finding algorithm

Software Tools	Peak Detection Method	Peak criteria	Peaks ranked by	Tag shift	FDR	User inputs	Differential peak calling?
MACS	Peak-finding	Local region Poisson P value	P value	Estimated from high quality peak pairs	1.None 2. $\frac{\#\ control}{\#ChIP}$	P-value threshold, tag length, mfold for shift estimate	Yes
PeakSeq	Peak-finding	Local region binomial P value	q value	Input tag extension length	Poisson 1.background assumption 2. From binomial for sample plus control	Target FDR	No
SPP	Peak-finding	Poisson P value (paired peaks only)	P value	Maximal strand cross-correlation	1.Monte Carlo simulation 2. $\frac{\#\ control}{\#ChIP}$	Ratio to background	Yes
SICER	Peak-finding	P-value from random background mode, enrichment relative to control	q-value	Input tag extension length	1.None 2. From Poisson p-values	Window length, gap size, FDR (with control) or E-value (no control)	Yes
GPS	Probabilistic binding	User defined threshold based on Kullback-Leibler divergence	p-value	None	None	P-value threshold	No
MACE	Peak-pairing	Chebyshev Inequality ranking	Pseudo p-value	None	None	Genome intervals	No

3.1.1.3 Differential binding Analysis tools

Quantitative differential binding analysis be can be performed by comparing the read counts (e.g. DBChIP[167], DEseq[168], edgeR[169,170]) or read densities (MAnorm[171]) in peak regions between conditions. This approach provides a statistical assessment (p-value or q-value) of differential binding based on read-enrichment fold changes across conditions. Given the wide suite of tools available for differential binding analysis, selecting the right algorithm is largely determined by the datasets used as well as the biological question at hand. To assist in selecting the proper tool to use, based on the dataset, Steinhauser and colleagues published an informative decision tree[172] (**Figure 3.5**). For the purposes of the analysis in this study, four differential binding tools are recommended (**Figure 3-5, Table 3-5**).

3.1.1.1.1 DBChIP

In the first step of analysis, DBChIP merges the lists of called peaks, obtained from external software such as MACS, from multiple conditions into a single consensus set. Called peaks are clustered using agglomerative hierarchical clustering with centroid linkage. The clustering approach ultimately begins by considering each object as a single-element cluster using linkage function, and at each iterative step two clusters that are similar are combined into larger clusters. The genomic positions of the consensus sites are then calculated, and weighted, as the average of the predicted or called peaks within each cluster. To detect differential binding sites, DBChIP tests a null hypothesis of non-differential binding at each consensus site. These tests are conducted through a generalized linear model with negative binomial distribution in order to account for the over-dispersion amongst samples.

In the event that biological replicates are available, a Negative Binomial distribution is estimated by edgeR. This results in p-value, fold change estimates and user-specified FDR thresholds between samples for each site[167]. Although, it is possible to detect differential binding in the absence of biological replicates, the assumption that the background reads across ChIP samples is comparable can be incorporated into hypothesis testing. However, this is unlikely in most cases and it is recommended that DBChIP be utilized for samples with matched or biological controls.

3.1.1.1.2 ChIPComp

ChIPComp, developed by the Wu lab, extracts ChIP-Seq signals from different datasets by estimating and removing biological and technical artefacts through a generalized linear model with Poisson distribution [173]. ChIPComp's normalization approach assumes that the samples being compared have a non-trivial number of consensus peaks and that there are no global binding differences within the consensus peaks across all samples. Once consensus peaks are identified, ChIPComp performs the Wald's test[174] on log ChIP-Seq signals based on log read counts. The resulting p-values are then used as input for probability calculations of differential binding using a Bayesian approach[173].

3.1.1.1.3 MMDiff

Unlike most differential binding tools which use read count enrichment strategies to detect differential peaks, MMDiff uses a kernel-based non-parametric strategy that focuses on changes in peak profile between samples[175]. This approach is solely dependent on the spatial structure and features of the ChIP-peak, which allows for MMDiff to detect highly localized changes that alter the shape of the peak. Furthermore, this approach makes MMDiff highly robust towards normalization effects and independent of the signal-to-noise ratio.

In MMDiff, each peak is treated as a *distribution* over a finite space, determined by the starting points of the reads within peaks. As these distributions are highly variable across the genome, the MMDiff developers utilize a machine learning kernel-based non-parametric test known the maximum mean discrepancy (MMD)[176], to predict differential binding from the observed peak distributions. The MMD is defined as the largest difference in expectations of functions in the unit ball of reproducing kernel Hilbert space (RKHS)[176]. With this approach, peak distributions at the same loci in different biological replicates will be classified as more different than expected.

To mitigate the effect of biological variability using this approach, for each peak, the number of reads mapping to that peak is averaged across all samples. Peaks are then binned based on the averaged read counts per peak. In each bin, the probability of observing an MMD value between the biological replicates in said bin is calculated. This probability is expected to be at least as large as the MMD value for a single peak between conditions. Obtained p-values are then corrected using the Benjamini-Hochberg method[165] for multiple significance testing.

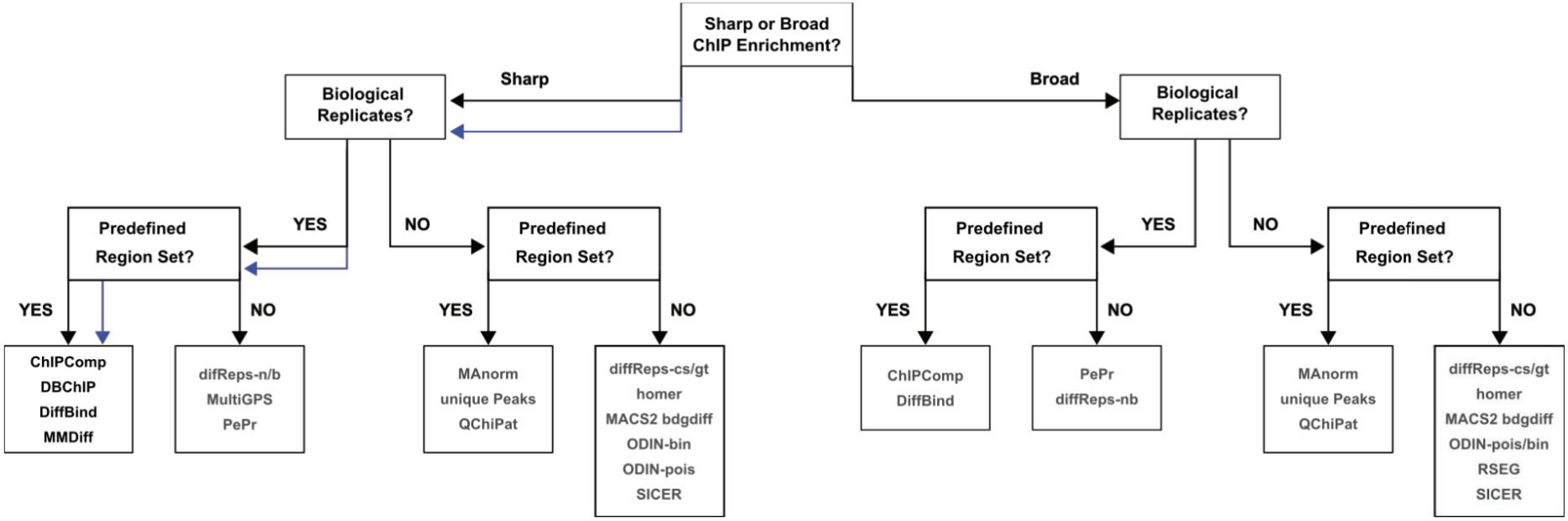


Figure 3-4: Decision tree of available differential binding analysis tools [172]. *For this study, Sharp Chip Enrichment with Biological Replicates within a Predefined Region set led to four potential differential analysis tools : ChIPComp, DBChIP, DiffBind and MMDiff. Arrows describe sequence of analysis events. Blue arrows indicate sequence used in this study ending with DiffBind differential DNA-binding factor analysis.*

Notably, the approach applied by MMDiff requires multiple biological replicates in order to reliably estimate the variance between samples. Although MMDiff is able to detect changes in homotypic binding events, it has been recommended as a complementary analysis tool to count-based approaches for exhaustive differential analysis[175].

3.1.1.1.4 DiffBind

Similar to ChIPComp, DiffBind creates a consensus set from called peaks, from which reads counted and subtracted from input/control datasets. Within DiffBind, RNA-Seq based normalization strategies, DeSeq, DeSeq2 or edgeR, can be carried out in order to identify differential peaks.

Table 3-3: Differential Binding Tools

Tool	Input	Prior Peak calling required?	Normalization	Statistical Test	Peak type	Biological replicates required	Significance measure	Characteristics
ChIPComp	Reads (*bed) Peaks (*.bed)	Yes	Generalized linear model with Poisson distribution	Wald's test followed by probability calculation using a Bayesian approach	Sharp	Yes	Posterior probability	ChIPComp can handle experiment designs of arbitrary complexity. It calculates the probability calculations of differential binding using statistical methods.
DBChIP	Reads (*.bed) Peaks (*.bed)	Yes	Mean ratio strategy (DESeq)	Generalized linear model with negative binomial distribution	Sharp	Yes/No	FDR	DBChIP is specifically designed for ChIP-Seq samples of transcription factors; it can handle experiment designs of arbitrary complexity (not limited to two-condition comparisons)
DiffBind: edgeR	Reads (*.bam) Peaks (*.bed)	Yes	Negative binomial distribution 1. DESeq 2. DeSeq2, or 3. edgeR		Sharp/ Broad	Yes	p-value or FDR	The method incorporates information from all peaks to estimate the common dispersion parameter, leading to a robust behaviour even with the minimal level of replication DESeq generalizes edgeR by allowing an arbitrary mean-variance relationship and, thus, is more adaptive to different datasets
MMDiff	Reads (*.bed) Peaks (*.bed)	Yes	DESeq	Kernel-based non-parametric	Sharp	Yes	p-value	This analysis uses the spatial structure of the ChIP-peak to determine differential binding, this allows for MMDiff to detect highly localized changes that alter the shape of the peak. MMDiff requires multiple biological replicates per condition.

3.1.1.1.4.1 Trimmed Mean of M-values normalization (edgeR)

The Trimmed Mean of M-values (TMM) normalization approach[177] employed by edgeR uses raw sample data to estimate the appropriate scaling factors to be used in downstream analysis. TMM normalization uses an empirical strategy that equates the overall binding affinities of peaks between samples under the assumption that the majority of peaks are not differentially bound and that the total read count is dependent on a few highly enriched loci. Thus, binding affinity log-fold changes can be reported as:

$$M_g(j,r) = \log_2 {}^{G_{gj}}\!/_{D_j} - \log_2 {}^{K_{gr}}\!/_{D_r}$$

and absolute intensity are reported as:

$$A_g(k,r) = \frac{1}{2}\left(\log_2\left({}^{G_{gj}}\!/_{D_j}\right) + \log_2\left({}^{K_{gr}}\!/_{D_r}\right)\right)$$

where K_{gl} = observed count for samples, g or r , in sample library, j

G= number of peaks

N= number of samples in the experiment

C_j= normalization factor of sample library, j

D_j= total number of reads for sample, j $\sum_g^G K_{gj}$

The trimmed mean is the average after removing the upper and lower x% of the data, which is doubly trimmed by the log-fold changes of $M_g(j,r)$ and $A_g(k,r)$. By default, $M_g(j,r)$ and $A_g(k,r)$ log fold changes are trimmed by 30% and 5%, respectively.

The weighted mean of M_g is calculated as follows:

$$TMM(j,r) = \frac{\sum_{g\in G*} w_g(j,r) M_g(j,r)}{\sum_{g\in G*} w_g(j,r)}$$

where $G*$ represents the set of loci with valid and untrimmed M_g and A_g

w_g= weighted mean

The inverse variance approximation for M_g is calculated using the delta method, which is then used to weight the average;

$$w_g(j,r) = \left(\frac{D_j - K_{gj}}{D_j K_{gj}} + \frac{D_r - K_{gr}}{D_r K_{gr}}\right)^{-1}$$

Thus, the correction factor can be obtained by

$$C_j = 2^{TMM(j,r)}$$

where all factors should multiple to one

$$C_j = \frac{exp\left\{\frac{1}{N}\sum_{l=1}^{N} \log\left(C_j\right)\right\}}{C_j}$$

These normalization factors across multiple samples can be calculated by selecting one or more samples as a reference followed by the calculation of the TMM factor for each of non-reference sample/s. The calculated TMM factors can then be built into statistical models such as the Poisson distribution to test for differential binding in samples as compared to reference sample/s. When the Poisson distribution model is used to detect differential binding, the observed library size is adjusted by a generalized linear model from a full library size to an effective library size. A full library size represents the total number of reads within BAM files while the effective library size is described as the number of reads mapped within peaks.

3.1.1.1.4.2 Relative Log Expression Normalization (DeSeq2)

DeSeq2 uses a relative log expression (RLE) normalization[168] approach that employs a geometric mean of raw read counts of the same locus in different samples. This strategy assumes that the read counts at a specific locus are proportional to DBP enrichment and sequencing depth. Similar to the TMM approach, by assuming that most loci are not differentially enriched, the median ratio C_j for a given sample is used as a correction factor for all read counts.

$$\left(\prod_{l=1}^{N} K_{gl}\right)^{1/N}$$

where K_{gl} = observed count for peak, g in sample library, j

g= number of peaks

N= number of samples in the experiment

C_j= normalization factor of sample, j

All samples in the experiment are then centred to the reference sample;

$$\frac{K_{gl}}{\left(\prod_{l=1}^{N} K_{gl}\right)^{1/N}}$$

The size or correction factor estimate C_j, is the median of ratios or geometric mean of the j-th samples read counts to those of a reference or pseudo-reference sample;

$$C_j = median_g \left\{ \frac{K_{gl}}{\left(\prod_{l=1}^{N} K_{gl}\right)^{1/N}} \right\}$$

where size factor estimates must multiple to one, similar to the TMM method, to obtain normalized counts:

$$C_j = \frac{exp\left\{\frac{1}{N}\sum_{l=1}^{N} \log\left(C_j\right)\right\}}{C_j}$$

3.1.1.4.1 Normalization for differential ChIP-Seq analysis

3.1.1.4.1.1 Relative level-difference (de novo normalization)

Differential binding analysis involves quantitatively determining similarities and differences between samples using peak intensity. This approach requires read normalization between samples. A simple normalization approach is to scale reads using the total read number within the whole genome or within background regions. In this approach reads are scaled using a constant factor or using a locally weighted regression (LOESS)[178]. This assumes that differences in the mapped reads between samples are small enough to be compared to the total read number and the genome-wide read count distributions have equal means and variances across samples. These assumptions may not be valid in a majority of cases where the signal-to-noise ratio between samples differs. Thus, peak calling algorithms using this scaling method need to consider different signal-to-noise ratios. For example, MANorm uses this scaling method using a robust linear regression that is based on the MA plot[171]. An MA plot is a scatter plot transformation of the M (log ratio) and A (mean average) scales between samples. ChIPComp, measures genomic background using the control sample and conducts a quantitative comparison of multiple ChIP samples which can also consider multiple-factor experimental designs [173].

3.1.1.4.1.2 Absolute-level difference (spike-in analysis)

Recently, the spike-in approach was described to normalize ChIP-Seq experiments in a computational-independent manner. In this approach, a small amount of chromatin from the genome of a different species with close enough homology that the DBP of interest shares conserved epitopes is added to experimental chromatin[179]. The spike addition into the experimental chromatin which will be ChIP-sequenced is typically added before immunoprecipitation and functions as an internal control for each sample. As the number of reads obtained from the internal reference chromatin is the same across all tested samples, the spike can then be used as an internal control for read normalization. This approach allows for the detection of global differences in DBP enrichment to the genome of interest at an absolute level[180]. Spike-in analysis is particularly useful in cases where genome-wide peak distributions change drastically i.e. knockdown or stimulated samples.

3.1.1.4.2 Signal-to-noise ratio

The signal-to-noise (S/N) ratio is determined by the number and intensity of peaks identified in each ChIP sample. ENCODE uses two metrics to determine S/N; fraction of reads in peaks (FRiP) and cross-correlation profiles (CCPs).

3.1.1.4.2.1 Fraction of reads in peaks

The fraction of reads in peaks (FRIP), positivity correlates with the number of peaks and the peak intensity of called peaks. The FRiP value is a simple ratio between the number of peaks (N_{peak}) identified and non-redundant DNA fragments in the library (N_{nonred}), calculated as;

$$FRiP = \frac{N_{peak}}{N_{nonred}}$$

The FRiP value, can be used as a threshold to filter out ChIP samples where too few peaks detected. Notably, the FRiP value is largely dependent on the sequencing depth and peak calling parameters and as such is not an objective or exhaustive metric.

3.1.1.4.2.2 Cross-correlation profiles (CCPs)

The cross-correlation profiles (CCPs) metric measures the clustering of reads prior to peak calling and can be described by Pearson correlations. The Pearson correlation coefficient of mapped read counts at each genomic position can be computed for each replicate or sample. Pearson cross-correlations between map read densities in both the positive and negative strands are plotted on the y-axis with strand shifts on the x-axis. Ideally, a test sample should have high S/N ratio i.e. high fragment length (C_{frag}) and low read length (C_{read}). Using the CCP approach, two quantitative metrics recommended by ENCODE can be scored;

1. Normalized strand coefficient, $NSC = \frac{C_{frag}}{C_{read}}$
2. Relative strand correlation, $RSC = \frac{C_{frag} - C_{min}}{C_{read} - C_{min}}$

where, C_{min} = minimum cross-correlation observed

C_{frag} = fragment length

C_{read} = read length

NSC and RSC scores can be calculated using phantompeakqualtools[157] independently from peak calling. For sharp or point-source DBP peaks, ENCODE recommends an $NSC \geq 1.05$ and an $RSC \geq 0.8$. Input and negative control samples on the other hand should have low NSC and RSC scores. A large S/N ratio suggests that there is a high number of enriched regions within the genome however, there is no guarantee that the identified peaks are "true" binding sites. This means samples with high numbers of false positives may also have high a S/N. Thus, this approach has not been widely applied and is currently limited to a few species.

3.1.1.2.2 Detection of motif binding sites with known PWMs

Following the fulfilment of the main aim of this study to identify promoter-associated (PA) genomic loci with lower *CTCF* enrichment in CRC from ChIP-Seq data, the canonical *CTCF* motif position weight matrix (PWM), MA0139.1[181], can be used to scan these regions for *"true"* motif binding sites. The detection and annotation of motif binding sites with known PWMs can be conducted using several databases and or tools described elsewhere.

In general, motif algorithms count the number of occurrences of each oligonucleotide of a given length in the ChIP'ed dataset which are compared with the number of occurrences that would occur by chance based on a background model. Background models are typically estimated form the composition of oligonucleotides within the ChIP'ed dataset. *"True"* binding sites typically have high probabilistic scores, i.e. p-value, against the PWM of interest, while alternative or background sequences will have low PWM p-values. Known motif finding requires a p-value cut-off score to threshold matched PWMs, the determination of which is inversely proportional to the number of "true" binding sites with strong binding affinity.

The p-value is represented as:

$$P(M_c, N_{R,c})$$

Where,

M_c = the set of motif sequences with cut-off, c.

$N_{R,c}$ = the number of k-mer sequences, $\{A_{s_1}, \ldots\ldots A_{s_m}\}$, with a PWM score $s_i > c$ within a the set of regions, R.

The p-value is defined as the probability to observed at least the same number, $N_{R,c}$ of motif instances with a cut-off, c, in a random sequence with a total length equal the total length of sequences in a set of regions, R[182]. Notably, the occurrence of clustered binding motif sites, as observed in *cis*-regulatory modules (CRMS), leads to their over-representation. However, most motif matching algorithms such as HOMER and FIMO[183], have user-specified cut-offs.

3.1.1.4 Integrative chromatin signature analysis using genomic platforms

The PA-LCe sites discovered by this pipeline, were analysed and annotated using several web-based annotation/visualization tools described in **Table 3-4**.

Table 3-4: Web-based annotation tools and browsers

Web-based Tool	Data analysed
UCSC genome browser	ChIP-Seq enhancer marks
	ChIP-Seq promoter marks
	COSMIC regions
	CpG Islands
	dbSNP
	DNAse I Hypersensitivity peak clusters
	GeneHancer
	GNF Expression Atlas 2
	gnomAD v3
	GTEx
	lincRNA RNA-Seq reads
	NHGRI-EB GWAS catalogue
	PolyA Transcript Annotation
	Protein Interactions
	TCGA Pan-Cancer (COAD & READ)
	Transcription Factor clusters
	Vertebrate Conservation
PrESSTo	FANTOM5 Human Promoters
	FANTOM5 Human Enhancers
	GTEx v6
GTex	RNA-Seq
	eQTL
	FANTOM5 CAGE
	PCAWG
ZENBU genome browser	FANTOM5 CAGE
PennState 3D genome browser	Hi-C
	ChIA-PET

3.2. PA-LCe discovery pipeline development

With the above-mentioned considerations for ChIP-Seq analysis, several analysis pipelines have been developed largely in a context-specific manner. In this study, we sought to identify promoter regions with lower *CTCF* enrichment from ENCODE *CTCF* ChIP-Seq data, in CRC cell lines as compared to normal colonic cells. This required a robust discovery pipeline that adhered to the ENCODE standards and practises as well as addresses the research question at hand.

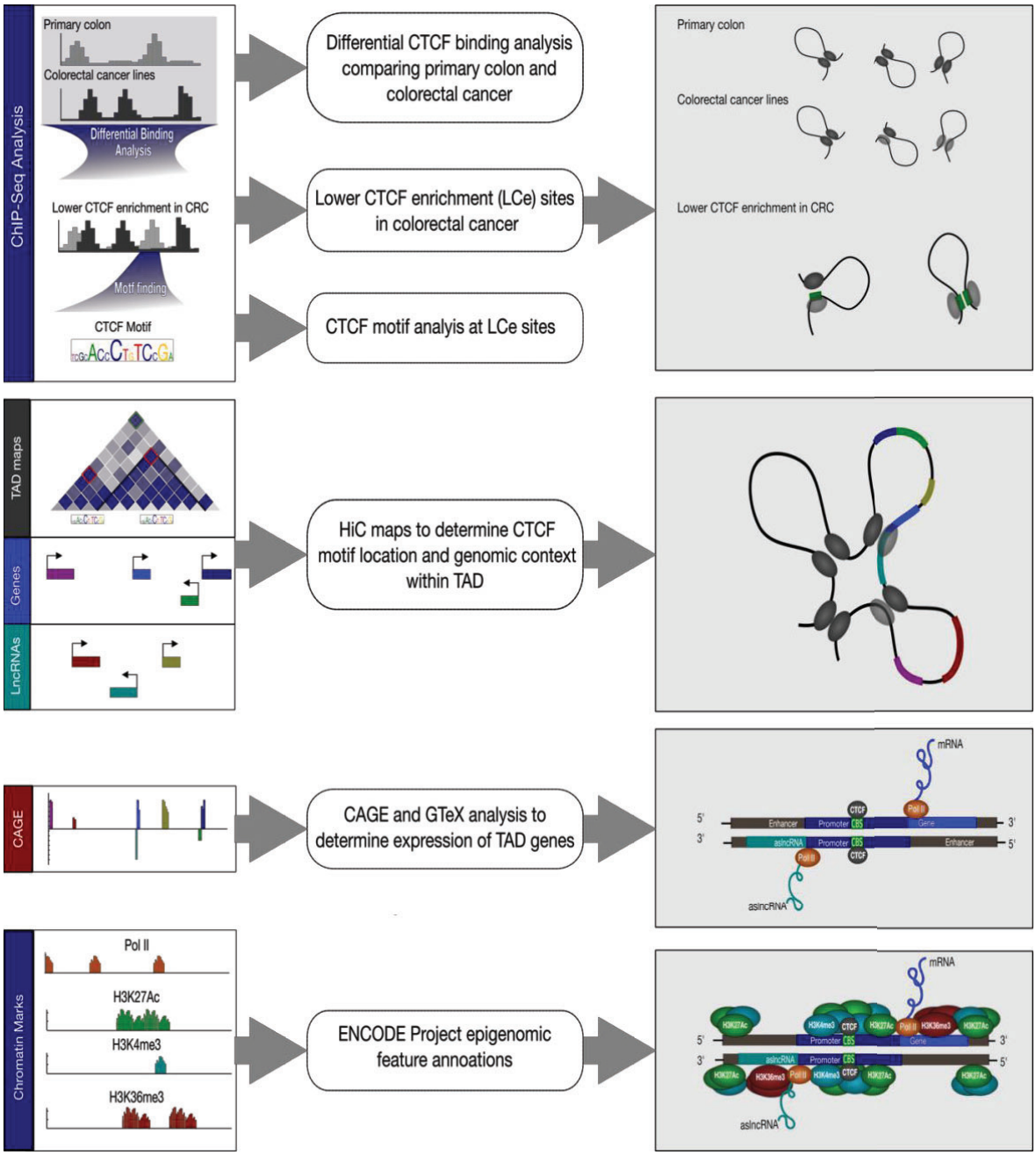


Figure legend in page 82

Figure 3-5: PA-LCe site discovery pipeline. *The pipeline begins with a customized bioinformatic ChIP-Seq analysis workflow to identify Promoter Associated Lower-CTCF Enrichment sites (PA-LCes), containing the canonical MA0139.1 CTCF motif, in colorectal cancer lines as compared to wild type sigmoid colonic cells. Chromatin conformation capture (HiC) maps at the identified PA-LCes are then analysed to determine the chromatin contacts, loops and TADS at these genomic loci. Genomic loci analysis is then used to determine and annotate proximal gene and lncRNAs. Following which, CAGE and GTex database analysis is then used to determine the transcriptional status of annotated genes and lncRNAs within these genomic regions in colorectal cancer samples. Finally annotated chromatin marks within these genes are used to determine the epigenetic context of PA-LCes containing genomic loci.*

To do this, the PA-LCe site discovery pipeline must fulfil the criteria listed in **Table 2.4**. The PA-LCe CBS discovery pipeline used in this study is described in this section is shown on **Figure 3-5.**

3.2.1 Dataset selection

All ENCODE ChIP-Seq datasets in this study were downloaded from the NCBI SRA[184]. As the datasets in this study were obtained from multiple studies, in an attempt to reduce batch effects selected datasets contained matched input controls. Of the publicly available datasets available for each cell or tissue type were selected using the specific criteria (**Table 3.5**).

3.2.2 ChIP-Seq Analysis methods

3.2.2.1 Read processing

3.2.2.1.1 FASTQCr: Quality control on raw reads

Raw FASTQ files were quality checked using FASTQC-0.11.3to determine read quality metrics. FASTQC-0.11.3. The dataset criteria for subsequent analysis are listed in **Table 3.5**. Basic read information, including total number of reads, total unique reads, read length were determined from raw FASTQ files using a custom bash script.

Table 3-5: Dataset selection criteria and quality metrics

Species	Human
Read type	SE or PE
Sequencing depth	>10 million reads
Controls	Required
Replicate availability	Optional
NRF Score	0.5≤ NRF ≥0.8
GC content (%)	>40%

3.2.2.1.1 Read mapping with bowtie2

ChIP-Seq FASTQ files were aligned to GRCh38 human reference genome using default parameters in bowtie2. Default parameters in bowtie2 are as follows:

Input files: FASTQ

Read trimming: 0

Phred scores: Phred+33 quality

Reporting: report best alignments with MAPQ

End-to-end alignment: entire read must align with no clipping

-D 15 : give up after 15 failed extends in arrow

-R 2 : for reads with repetitive seeds, 2 sets of seeds attempted

-N 0 : max number of mismatches in seed alignment= 0

-L 22: length of seed substrings = 22 bp

-i S,1,1.15: interval between seed substrings with respect to read length

3.2.2.1.2 Post-alignment processing

The aligned reads, in SAM file format, were sorted and converted into BAM files using samtools. Aligned and unique reads with MAPQ ≥ 30, were extracted using samtools from BAM files. In most well prepared ChIP-Seq experiments, most of the duplicated reads are likely to be “true” duplicates suggesting high levels of enrichment. Notably, the primary aim of this study is to identify differential, however marginal, *CTCF* enrichment sites. Thus, to avoid

decreasing signal strength and to identify regions with differential signal enrichment, duplicate reads in this study were not removed. Instead, duplicate reads were marked using Picard's MarkDuplicates. The filtered reads were also indexed using samtools to produce .bai files.

3.2.2.1.3 Quality control on processed reads

Quality metrics for unprocessed, and processed BAM files were reviewed using FASTQCr. The aligned and processed reads underwent quality control checks and were visualized using with FASTQC, IGV.

3.2.2.2 Peak Calling

Narrow peaks were called according to ENCODE ChIP-Seq guidelines using MACS2 with a p-value cutoff for peak detection set to 0.01 with the effective genome size set to "human" (2.7e9). Blacklisted regions were subtracted from narrow peaks that were called. Peaks in variable alternate scaffold loci and chrM (/chrM/d;/random/d;/alt/d;/chrUn/d) were also removed post peak calling using bedtools as well as the blacklist GRCh38 regions.

3.2.2.3 ChIP-QC

Quality control metrics on processed BAM files and corresponding narrowPeak files were computed using the ChIP-QC package in R. The main QC metrics on reads include; total number of reads in BAM file for each sample, the percentage of these reads that were successfully aligned to GRCh38 and the percentage of reads with a MAPQ ≥ 30. Using processed peak calls, ChIP-QC was used to compute peak-based metrics peak-based metrics, such as FRIP, peak profiles, and clustering.

3.2.2.3 Differential binding analysis with DiffBind

3.2.2.3.1 Dataset classifications

In this study, ENCODE datasets from primary cell/tissue samples and cell lines were used. None of the datasets used in this study contained biological replicates, thus for the purposes of this analysis, biological replicates were defined as datasets from the same sample type. For the colonic datasets, the primary condition had two replicates (SColon 37 and 54) and CRC lines had 3 replicates (Caco2, DLD1 and HCT116). In order to verify the robustness of

the PA-LCe discovery pipeline in determining tissue-specific PA-LCes, we implemented the pipeline on a leukaemia dataset. This dataset, also obtained from ENCODE, included CTCF-ChIP-Seq datasets of CD14+ monocytes and GM12878 were treated as biological replicates for the primary cells and lymphoblastic leukaemia cell line K562 was treated as the cancer line.

3.2.2.3.2 Raw Peak Correlations

Pearson correlations and Principal Component analysis on called peaks were conducted using the dba.plotheatmap and dba.plotPCA functions in DiffBind, respectively.

3.2.2.3.3 Normalized Peak Correlations

A binding matrix containing normalized scores based on the read counts per sample at every potential binding site was generated using the DiffBind dba.count function and used for subsequent differential analysis. Reads in CTCF binding site intervals were counted, scored and normalized using DeSeq2 in DiffBind[185,186]. DeSeq2 uses a read count based normalization strategy where the raw number of reads in the control sample is subtracted prior to the library size computation. The total number of reads in each dataset was used as the library size (bFullLibrarySize = TRUE). Pearson correlations and Principal Component analysis on normalized peaks was also conducted using the dba.plotheatmap and dba.plotPCA functions in DiffBind respectively.

3.2.2.3.5 Differential binding analysis

Read count correlations between datasets and conditions were calculated using a p-value threshold of 0.05. Differential binding analysis between the wild-type and cancer conditions was conducted with DESEQ2 using the full library size with an FDR < 0.05. All differentially binding conditions were saved as bed files for peaks with increased, decreased and non-differentially bound CTCF peaks.

Prior to differential binding analysis, contrasts were established to classify samples into condition groups i.e. primary and cancer conditions using the dba.contrast function. Differential binding analysis was executed using the core DiffBind function dba.analyze using DESeq2. This resulted in the assignment of p-values and FDR scores, to each candidate

binding site in order to indicate the confidence peaks identified were differentially bound. Visualizations of the differential binding analysis report including; Pearson correlation heatmaps, PCA, MA, scatter, violin and boxplots were generated in DiffBind.

Differential binding analysis data was extracted into data frames in RStudio as follows:

1. Reduced Affinity binding peaks (db.loss),

2. Increased Affinity binding peaks (db.gain)

3. Stable differentially bound binding peaks (db.notdb)

These data frames were also converted into bed and fasta files using bedr[187] for subsequent analysis.

3.2.2.4 Differential Peak annotations

3.2.2.4.1 Chip annotations

Differential binding data frames resulting from DiffBind were annotated using the ChIPpeakAnno package. A tag matrix for each data frame was generated and used to create tagHeatmaps and coverage plots of differentially bound sites 10kb around hg38 promoter regions. Other annotations and visualizations included; upset plots and Venn pie charts.

3.2.2.5 Motif discovery using HOMER

De novo motif discovery on bed file outputs was conducted using the HOMER[188] annotatePeaks tool in order to determine enriched motifs in differentially bound datasets. Motif discovery with HOMER was conducted on both the bed file and fasta outputs from **Section 2.2.2.4** to annotate differentially bound peaks with the canonical CTCF (MA0139.1[181]) motif.

3.2.2.5 Promoter-associated lower CTCF-enrichment motifs

The binding of CTCF to gene promoters has been shown to be a requirement for the adequate transcription of several tumor suppressors. Thus, LCe sites containing a canonical MA0139.1 motif within promoter regions (PA-LCe sites) were extracted from annotated CTCF motif feature files.

Chapter 4 : PA-LCe sites in CRC are associated with enhancer-derived antisense long non-coding RNAs

4.1 Introduction

Human cancers are fundamentally heterogeneous with distinct subtypes associated with differences in genetic, molecular, cellular, pathological and clinical presentations. However, most cancer-related mutations occur in regions of the genome outside of genes. However it difficult to determine which of these non-coding mutations have functional relevance in cancer and which are merely just noise. Although progressing, our limited understanding of how these non-coding regions regulate transcriptomic activity is still in its infancy. What has become increasing evident, is that the non-coding genome regulates the coding genome through multiple co-operative modes of action.

Transcription of the human genome is proximity-dependent requiring chromatin contacts between promoters and other genomic loci, such as enhancers regions, and transcriptional machinery. The repertoire of contacts chromatin is engaged in at any given time is regulated by the three dimensional structure and organization of the chromatin. Indeed, aberrant contacts particularly between enhancer regions and cancer-specific genes i.e. oncogenes and tumor suppressors, that have been documented as drivers of oncogenesis. Classical examples include the acquired contact between the *PDGFRA* proto-oncogene and enhancer in the neighbouring TAD which results in the activation of *PDGFRA* in *IDH*-mutant gliomas[16]. Similarly, in lymphoblastic leukaemia tumors, the proto-oncogene *Lmo2* is activated by the creation of new enhancer-promoter contacts[17]. In both these instances, the formation of ectopic cancer-specific enhancer-promoter contacts is facilitated by disruptions in *CTCF*-mediated chromatin insulation. Mutations in CTCF binding sites have been frequently observed in multiple cancer tumors. In CRC specifically, the mutational profiles of CBs have been documented by several studies [77–79]. Collectively, these studies have demonstrated dysregulated CTCF binding, either by somatic mutation or DNA methylation, as a putative oncogenic mechanism that results in cancer-specific chromatin contacts and thus altered transcriptional programs. Thus, CBSs with abrogated CTCF binding in cancer can be used as potential diagnostic and therapeutic markers. Promoter-resident CBSs functioning as “enhancer-docking sites” which regulate enhancer-promoter contacts, and transcriptional activation in a cancer and tissue-specific manner [88], have emerged as attractive targets.

In this study, we sought to identify and characterize promoter-associated CTCF binding sites with abrogated *CTCF* enrichment (PA-LCes) using ENCODE CTCF ChIP-Seq datasets. The full criterion of a PA-LCe site is in two parts:

1. Lower CTCF enrichment in the cancer dataset as compared to primary/wild-type cells.

2. The PA-LCe genomic region must be <1kb away from the closest annotated TSS in the hg38 human genome."

In order to address identify these PA-LCe sites, we sought to develop a straightforward and robust PA-LCe discovery pipeline with open-source tools (described in **Chapter 3**). Furthermore, we utilized various genetic and epigenetic platforms as well as studies to gain insight into functional relevance of the identified PA-LCe sites in cancer.

4.2 Methods

4.2.1. Datasets

All ChIP-Seq data in CRC sample set was downloaded from the NCBI SRA[184] using fastqdump. As the datasets in this study were obtained from multiple studies, in an attempt to reduce batch effects, the datasets selected had to contain matching input controls. ENCODE. *CTCF* ChIP-Seq datasets of primary sigmoid colon tissue and CRC cell lines were collected from NCBI SRA using fastq-dump.2.8.1. The full list of the selected datasets falling within these criteria can be found in (**Table 4-1**).

Table 4-1: CRC datasets used in this study

Dataset Name	SColon37	SColon54	HCT116	DLD-1	Caco-2
Organism	Human				
Tissue	Sigmoid Colon		Colon		
Cancer type	None		Carcinoma	Adreno-carcinoma	
Disease	Anoxia		-	Dukes' type C	
Cell Type	Muscularis propria		Epithelial		
Age	37	54	Adult		72
Gender	Male				
Data source	Michael Snyder (Standford)		Various labs		
Read type	Paired-end		Single		

4.2.2. Bioinformatic pipeline

The PA-LCe *CTCF* motif in CRC discovery pipeline developed (**Chapter 3**) conducted in this study is summarised in **Figure 4-1** and **Table 4.2**.

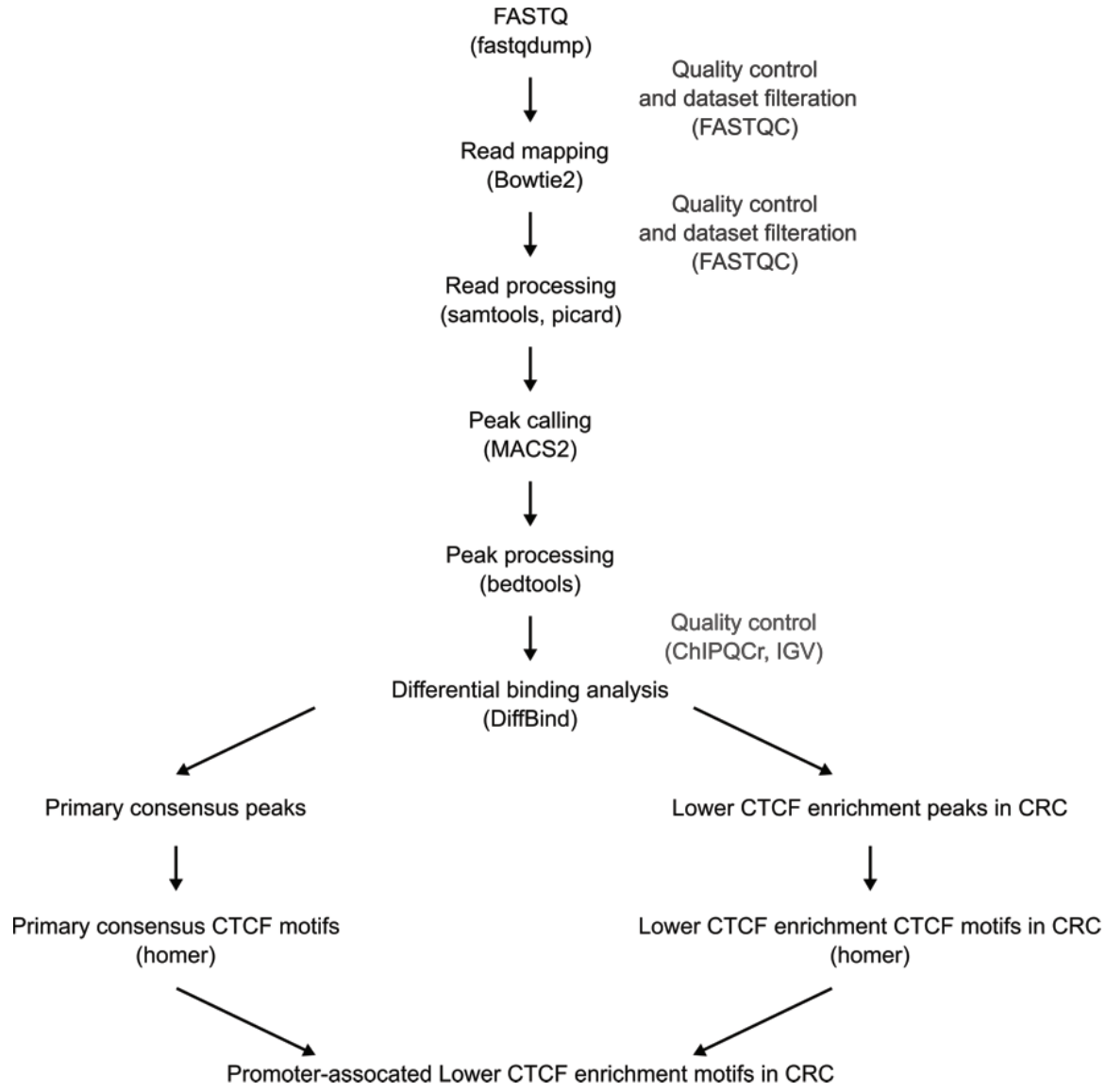


Figure 4-1: Differential ChIP-Seq analysis workflow and tools used to determine PA-LCe CTCF motifs in CRC*. The workflow subjects each fastq file to a quality control and filteration step using FASTQC. Quality checked reads in fastq files are then mapped to the human genome with Bowtie2 and processed using samtools and picard. Narrow peaks are called from processed reads using MACS2. Called peaks are processed using bedtools. Differential peak analysis between called colorectal cancer (CRC) and wild type peaks is performed using DiffBind after undergoing quality control with ChIPQCr and IGV. Differentially called peaks are classified into consensus peaks, lower and higher enrichment peaks within DiffBind. CTCF motif (MA0139.1) discovery is performed with homer on lower enrichment and primary consensus peaks which are then annotated to identify lower enrichment motifs in CRC cell lines as compared to primary colonic samples*

4.3 Results and Discussion

4.3.1. PA-LCe Discovery in CRC

4.3.1.1 Quality control: FASTQC

Basic sequence statistics (total reads, total number of unique reads, the percentage of unique reads, the most abundant sequence, its frequency and the percentage of distinct uniquely mapping reads, read length, maxRead sequences, count and score) were determined using a custom bash script.

The above-mentioned metrics were used to assess the datasets that would be used for subsequent analysis. All fastq files collected **4.1.1** underwent FASTQC analysis. The full list of FASTQC results is shown in **Section 5.2,** an example of ideal FASTQC results (*CTCF* HCT116), is shown in **Figure 4-1**. Samples not fulfilling the criteria in **Table 3-4** were removed from subsequent analysis.

Measure	Value
Filename	CRC_HCT116_CTCF_Test_3.1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	19158950
Sequences flagged as poor quality	0
Sequence length	36
%GC	45

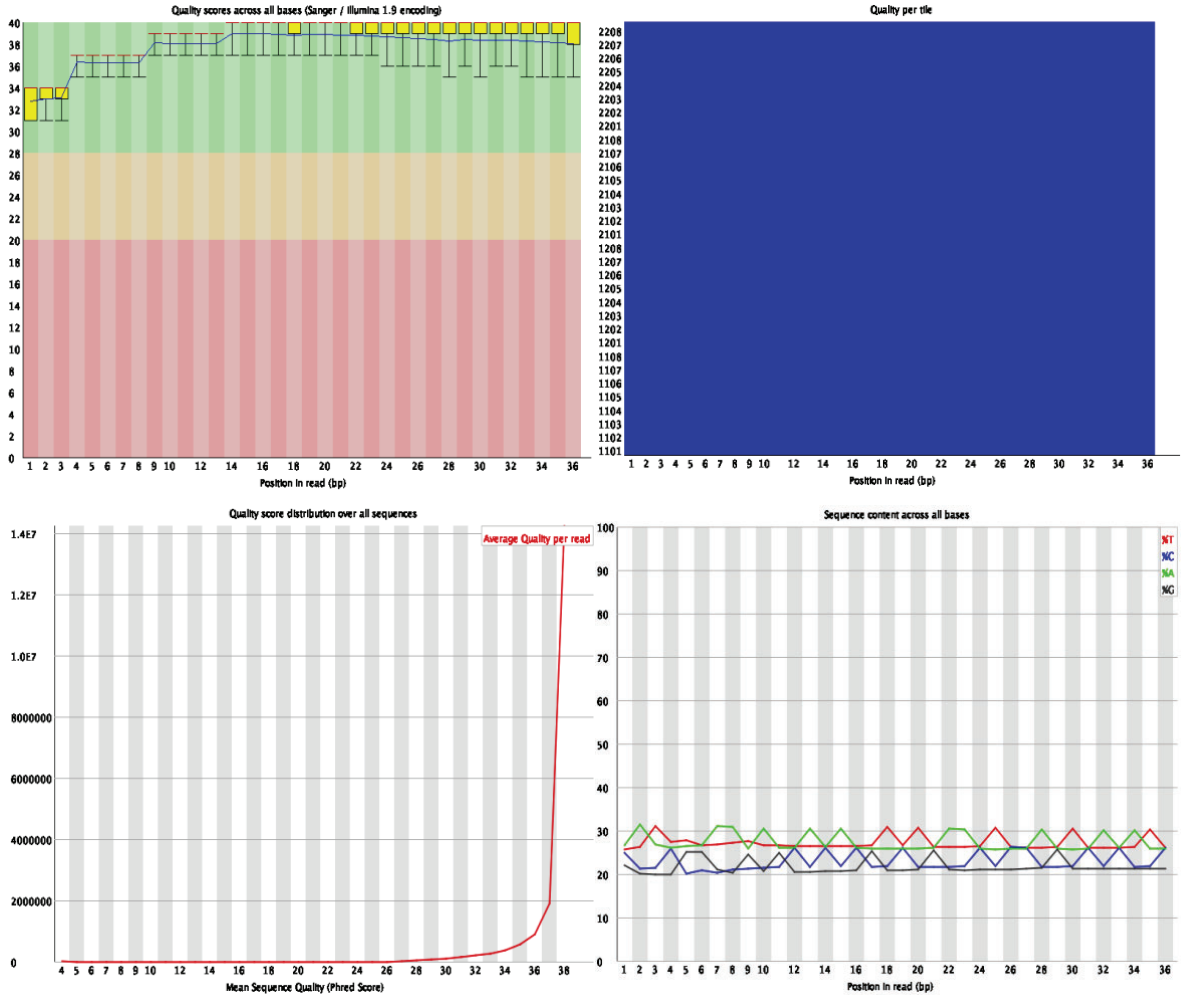
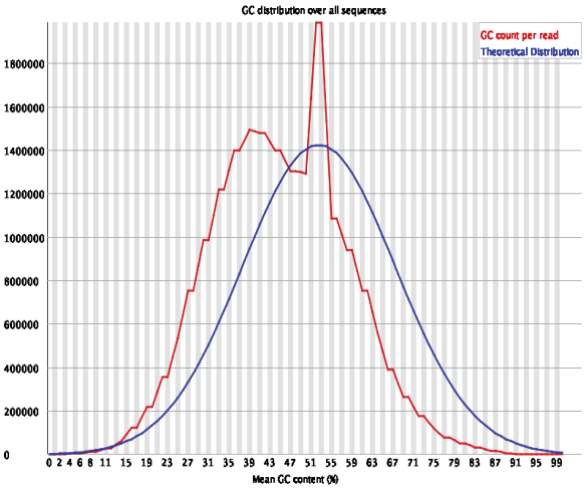
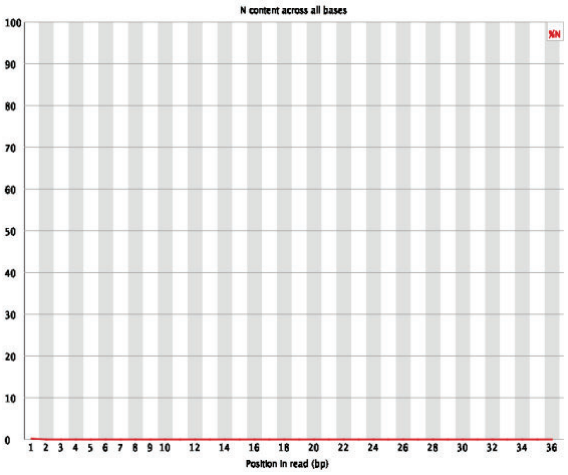
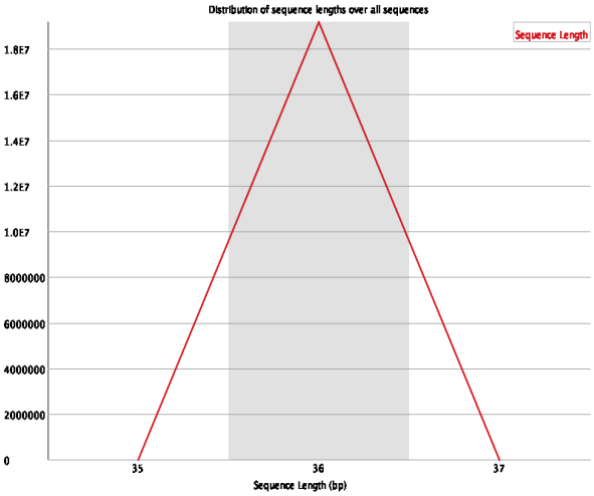
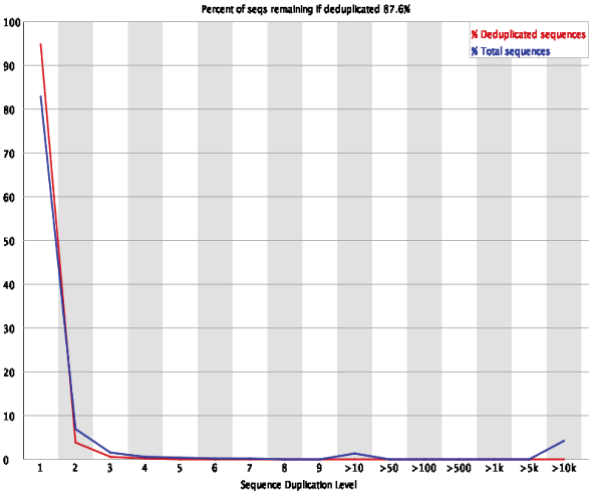


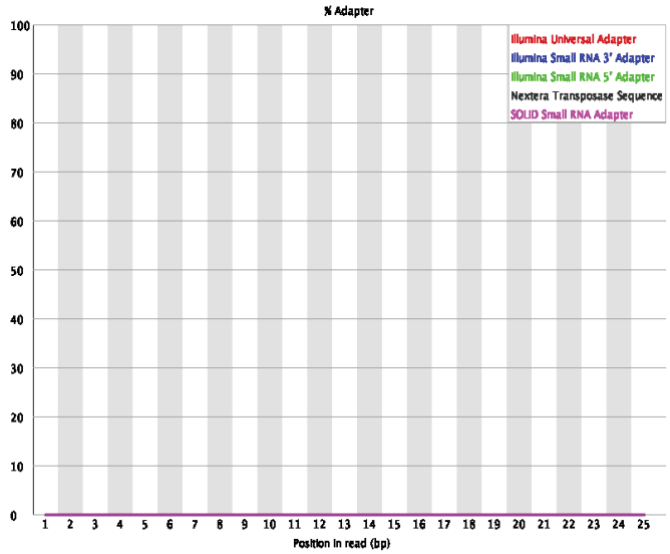
Figure 4-2: FASTQCr report of the CTCF HCT116 processed BAM*. Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, Per base N content, Sequence length distribution, Sequence duplication levels, Overrepresented sequences, Adapter Content. All FastQCr reports can be found in* ***Section 5.2****.*











Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATC	810157	4.228608561533904	TruSeq Adapter, Index 1 (100% over 36bp)
ATCGGAAGAGCACACGTCTGAACTCCAGTCACATCA	29842	0.15576010167571813	TruSeq Adapter, Index 1 (100% over 36bp)

Legend in page 91.

4.3.1.2 Reading mapping and processing

ChIP-Seq FASTQ files were aligned to GRCh38 human reference genome using default parameters in bowtie2 as described in **2.3.2.1**. Unique and mapped reads with MAPQ > 30 were extracted for subsequent analysis using samtools.

Duplicate reads were marked with picard MarkDuplicates. Resultant files were indexed using samtools. Filtered and unfiltered read files were analyzed using FASTQCr, the results can be found in **Section 5.2**. Raw and processed reads visualized on IGV (**Figure 4-3**).

4.3.1.3 Peak calling

Narrow peaks were called according to ENCODE's ChIP-Seq guidelines using MACS2 with a p-value cutoff for peak detection set to 0.01 and the effective genome size set to "human" (2.7e9). Blacklisted regions (**Table 6.3**) and variable alternate scaffold loci were subtracted from called narrow peaks as described in **2.3.2.2**. For visualized inspection of called peaks corresponding to mapped reads, BAM coverage, and read files along with corresponding peak calls from MACS2 were loaded onto IGV as seen in **Figure 4-3**.

4.3.2 ChIP-QC

Quality control metrics on processed BAM and corresponding narrowPeak files were computed using ChIP-QC in R. A summary of ChIP-QC results **Table 4-2**. The RelativeCC metric, which is calculated by comparing the maximum cross coverage peak to the cross coverage shift corresponding the read length for most datasets is greater than 1 indicating good enrichment and high quality datasets. Notably, the RelCC values for the primary dataset controls are almost 2-fold greater than that of the *CTCF*-ChIP'ed datasets. As seen in the code 4. Differential Peaks with DiffBind.Rmd,, the bAddCallerConsensus argument in ChIPQCr has been set to FALSE. If set to TRUE, ChIPQCr will generate a consensus peak set derived by merging all overlapping peaks in all provided peaksets, test and control alike, only keeping peaks that overlapping in at least two samples. This means, peaks found in only one library control will be excluded from subsequent differential binding analysis. For the purposes of this study this behaviour is not ideal has the removal of peaks found only within one library, specifically a control library, would be excluded and significantly reduce the number of differentially enriched sites obtained in this pipeline. Thus, to utilize all MACS2 called peaks within each dataset, the bAddCallerConsensus argument was set to FALSE. This argument leads ChIP-QC to assume the replicate numbers of these peaks are missing as observed in **Table 4-4**.

The density of positions with different pileup values in primary CTCF-ChIP'ed datasets was greater than 1.5 indicating higher enrichment whereas SSD values for all CRC datasets were below 1 indicating an overall lower enrichment as compared to primary datasets. These SSD values correspond to the variance in sequencing depth across the different datasets. The signal-to-noise ratios of all the ChIP'ed datasets were greater than 5% indicating successful enrichment with no peaks mapped to blacklisted regions. Overall, the datasets used had relatively acceptable quality metrics.

Figure 4-3: IGV visualization of the MYC locus in SColon37 and HCT116 datasets displaying ChIP-Seq peaks discovered by MACS2*. BAM coverage peaks (histograms) and BAM reads (grey bars) of SColon37 (blue) and HCT116 (red) datasets.*

Table 4-2: Summary of CRC dataset ChIP-Seq filtering and quality metrics produced by ChIP-QC

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%	RiBL%
CTCF_Caco2	Caco2	CTCF	Cancer	2	1075081	24	36	116	2.8	0.77	19	0
CTCF_DLD1	DLD1	CTCF	Cancer	1	2918169	30	36	174	3.6	0.8	15	0
CTCF_HCT116	HCT116	CTCF	Cancer	6	1143583	5.9	36	148	3.2	0.92	33	0
CTCF_SColon37	SColon37	CTCF	Primary	1	5616024	22	101	225	2.6	1.8	11	0
CTCF_SColon54	SColon54	CTCF	Primary	1	15385137	34	101	234	2.7	1.9	5.4	0
CTCF_SColon37_Control	NA	NA	NA	NA	4132437	3.4	101	205	5.6	0.88	NA	0
CTCF_Caco2_Control	NA	NA	NA	NA	756702	0.61	36	81	1.4	0.4	NA	0
CTCF_DLD1_Control	NA	NA	NA	NA	1311652	18	36	73	14	0.5	NA	0
CTCF_HCT116_Control	NA	NA	NA	NA	1430294	0.4	36	74	0.72	0.4	NA	0
CTCF_SColon54_Control	NA	NA	NA	NA	7845323	12	101	204	5	1.1	NA	0

The processed reads for each of the datasets in this study passed the MAPQ filter as all aligned reads were filtered for reads with MAPQ vales ≥ 30 (**Section 2.3.2.1.2**). As duplicated reads were not removed in this study, the percentage of duplicate reads in these datasets ranged from 0.41 to 30% with the highest duplication reported in both input and *CTCF*-ChIP'ed datasets.

Table 4-3: Number and percentage of mapped, duplicated and MapQ filter passing reads

ID	Tissue	Factor	Condition	Replicate	Unmapped	Mapped	Pass MapQ Filter and Dup	Total Dup%	Pass MapQ Filter%	Pass MapQ Filter and Dup%
CTCF_Caco2	Caco2	CTCF	Cancer	2	0	1075081	257310	24	100	24
CTCF_DLD1	DLD1	CTCF	Cancer	1	0	2918169	869217	30	100	30
CTCF_HCT116	HCT116	CTCF	Cancer	6	0	1143583	67118	5.9	100	5.9
CTCF_SColon37	SColon37	CTCF	Primary	1	0	5616024	1240004	22	100	22
CTCF_SColon54	SColon54	CTCF	Primary	1	0	15385137	5180840	34	100	34
CTCF_SColon37_Control	NA	NA	NA	NA	0	4132437	140625	3.4	100	3.4
CTCF_Caco2_Control	NA	NA	NA	NA	0	756702	4608	0.61	100	0.61
CTCF_DLD1_Control	NA	NA	NA	NA	0	1311652	229034	17	100	17
CTCF_HCT116_Control	NA	NA	NA	NA	0	1430294	5799	0.41	100	0.41
CTCF_SColon54_Control	NA	NA	NA	NA	0	7845323	949146	12	100	12

All cross-coverage plots (**Figure 4-3b**) peak at a shift-size greater than $1.5 \times read\ length$ and correlate to the fragment lengths reported in **Table 4.2**. As shown in **Figure 4-2c** the coverage histogram, demonstrating the distribution of pileup values at each base pair, shows the primary datasets contain more CTCF enrichment as compared to the CRC datasets. This suggests that the CRC peaks have lower read densities in called peaks as compared to the primary dataset. It is important to note that the higher CTCF-enrichment in the primary samples that are used as "controls" in this study may introduce bias in the analysis as prior to differential analysis, the data is skewed towards CRC lines containing peaks with lower read densities as compared to primary datasets.

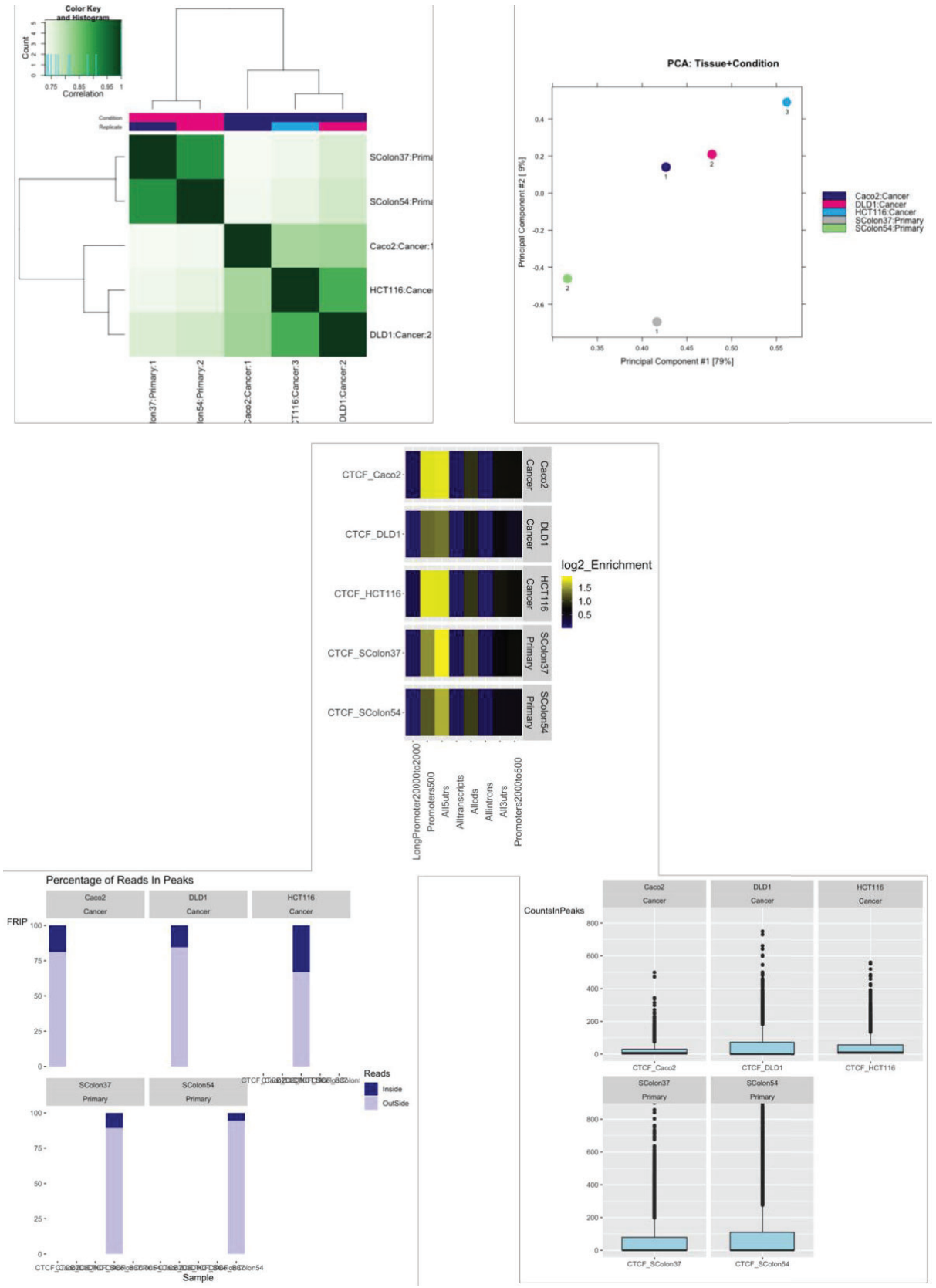


Figure legend in page 97

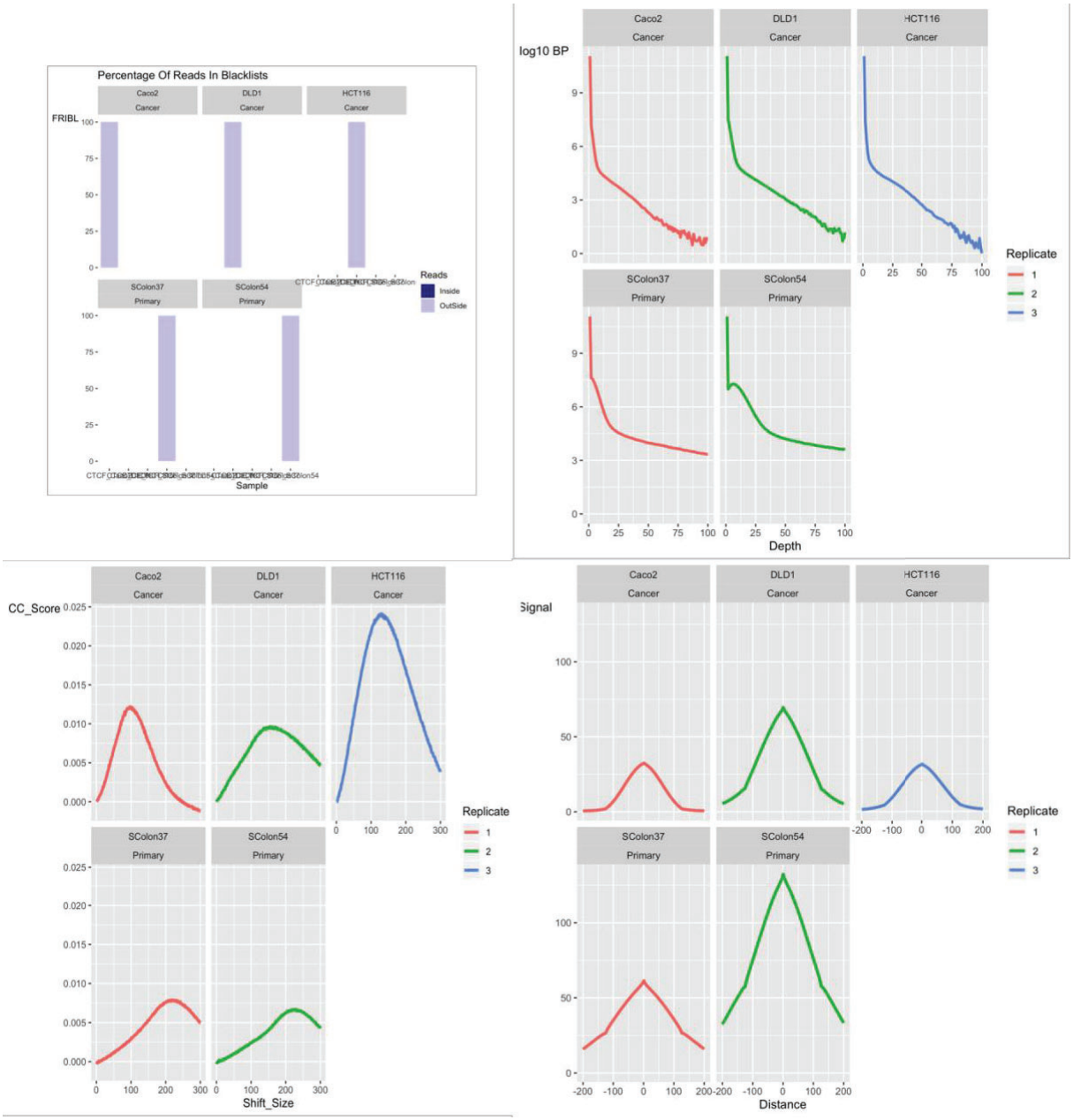


Figure 4-4: ChIPQC Results for CRC dataset. Top panel: *Pearson Correlation plot of peak sets and Principal component analysis of peak sets.* ***Centre panel:*** *Log2 enrichment of read counts at specific genomic annotations; Bar plot of the percentage number of reads in peak; Density plot of the number of read counts in peaks; and Bar plot of the percentage of reads in blacklist,* ***Bottom panel:*** *Log2 base pairs of genome at differing read lengths; Cross Coverage score after successive strand shifts; and Plot of the average signal profile across peaks*

The ChIP efficiencies reported by the fraction of mapped reads (FRIP) for the colonic peaksets was relatively higher for CRC (15-30%) as compared to primary colon (<15%) (**Figure 4-4**). This suggests primary colon peaksets contained higher background as compared to CRC lines. However, the primary colon peakset contained a greater number of read counts within

called peaks as compared to the CRC lines, corresponding with the global reduction of read densities in peak regions observed in **Figure 4-4**. Peak profiles for all datasets were all centred and peaked around the summit regions in **+/-**200bp windows with the highest signal enrichment observed in primary colonic datasets, as well as the DLD1 dataset (**Figure 4-4**). The raw colonic peaksets are distinctly clustered by the condition; primary and CRC datasets respectively, as displayed by the Pearson correlation heatmaps and PCA plots (**Figure 4-4**).

4.3.3 Normalization tests in the PA-LCe discovery pipeline

As the datasets used in this here were obtained from various labs within the ENCODE Consortium, it was important to normalize the called peaks to obtain differentially CTCF binding. As recommended by the decision tree published by Steinhauser and colleagues [172], we used DiffBind to extract differentially enriched peaks between the primary and CRC datasets. Within DiffBind, we performed the normalization strategies using DeSeq2 and edgeR with either the full (total number of reads within BAM files) or effective (number of reads mapped within peaks) library size on ChIP-Seq read counts (**Figure 4-5**). For this analysis the primary datasets were treated as replicates of a reference sample that the CRC replicates were compared to and the false discovery rate for all normalization conditions was <0.05.

The TMM normalization[177] strategy employed by edgeR assumes that the majority of peaks between the primary and cancer samples are not differentially bound. This strategy also assumes that the total read count is dependent on a few highly enriched loci. Similar to edgeR's TMM normalization approach, the relative log expression normalization[168] strategy used by DeSeq2 assumes that most loci are not differentially enriched but also extends the edgeR approach by subtracting the scaled input/background read counts from the overlapping peaks. Unlike, the TMM normalization approach however, DeSeq2 assumes that the read counts at a specific locus is proportional to the enrichment of the DBP as well as the sequencing depth associated with each library (reads in associated BAM file).

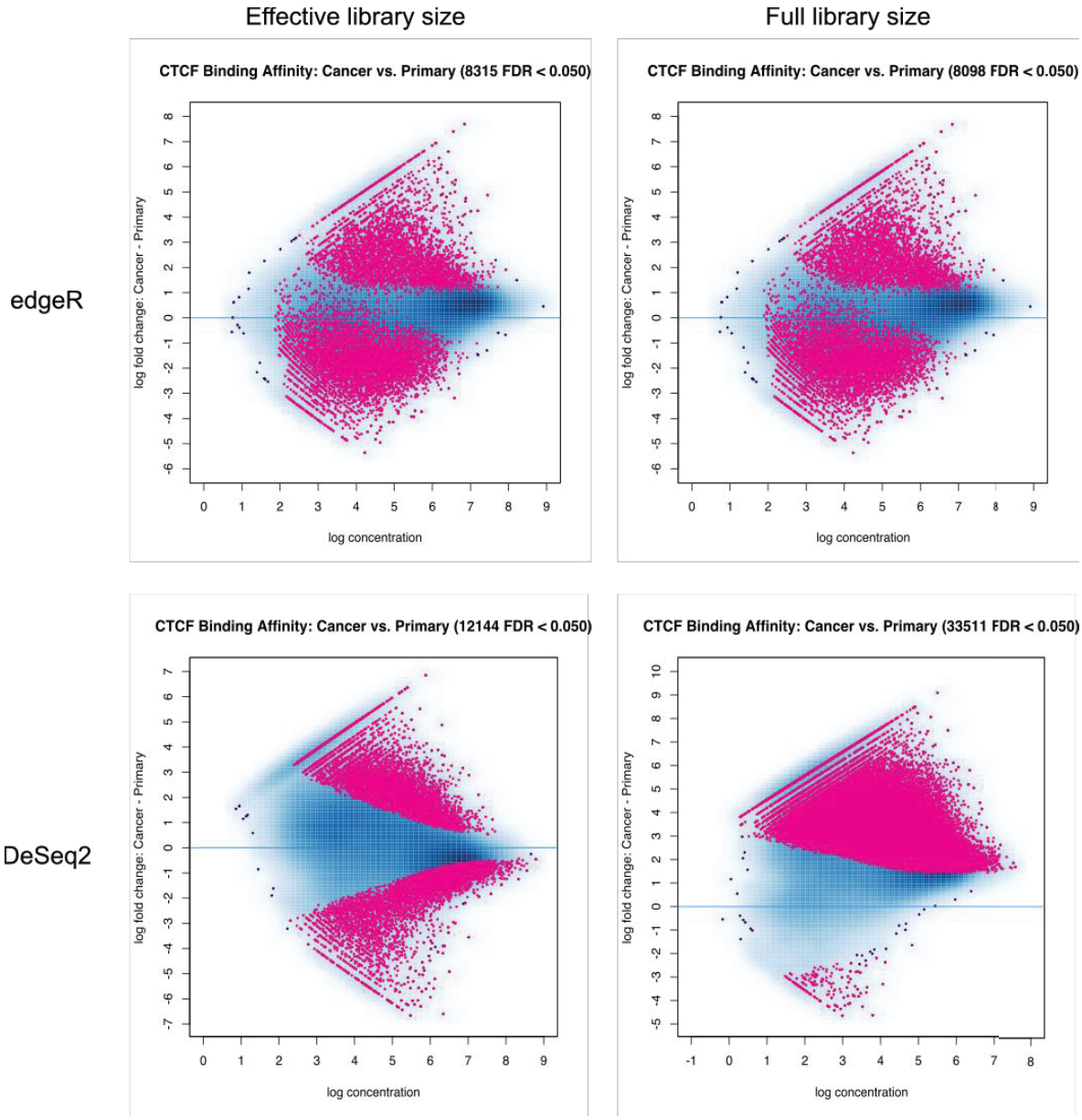


Figure 4-5: DiffBind normalization of CTCF ChIP-Seq read count data using edgeR and DeSeq2 with full and effective library sizes as displayed by MA plots. *Dataset comparing colorectal cancer cell lines and primary sigmoidal colon cells. Each point represents a binding site, with pink points representing sites identified as differentially bound in colorectal cancer cell lines as compared to primary sigmoid colon cells (FDR <=0.05). Blue points represent non-differentially bound sites between the two conditions. M-value on vertical axis and peak height A on the horizontal axis.*

Normalizing for library size, for both edgeR and DeSeq2 analysis methods, resulted in significant differences between the cancer (CRC) and primary datasets (**Figure 4-5**). In both the edgeR and DeSeq2 normalizations for both library size conditions, differentially enriched CTCF peaks were skewed towards sites with increased CTCF enrichment in CRC(**Figure 4-**

5). The highest absolute log fold differences observed in the DeSeq2 full library size normalization condition as compared to sites with reduced CTCF enrichment (**Figure 4-5**).

Biologically we expect a high number of differentially bound sites in CRC. Using edgeR we observed high numbers stable CTCF-enrichment peaks, 22 188 and 22 305 for effective and full library sizes, respectively (**Figure 4-5**). While DeSeq2 reported lower number of stably CTCF-enrichment peaks between conditions, 23 037 and 17 640 for the effective and full library sizes, respectively (**Figure 4-5**). Using edgeR analysed datasets we obtained 8 315 and 8 098 differentially bound peaks using the effective and full library size, respectively (**Figure 4-5**). With DeSeq2 we observed a higher number of differentially bound peaks as compared to edgeR, with 12 144 and 33 511 differentially bound peaks using the effective and full library size, respectively. Thus, MA plots using DeSeq2 using the full library size revealed a large number of differentially enriched CTCF peaks in CRC (**Figure 4-5**) with the data skewed towards increased CTCF enrichment in the CRC dataset as compared to the primary dataset. This observation fulfils the biological expectation of obtaining a high number of differentially bound sites in the CRC condition. Furthermore, DeSeq2 further extends this assumption to subtract scaled/control reads from all overlapping peaks within the full library providing higher confidence in the differentially enriched peaks identified We opted to use the DeSeq2 full library size normalization strategy for subsequent analysis.

4.4 PA-LCe discovery pipeline reveals lower CTCF enrichment at as-lncRNAs promoters in CRC

In this study, we hypothesized that sites with abrogated CTCF enrichment in CRC could be used as diagnostic and therapeutic markers for CRC. To this end, we developed a bioinformatic discovery pipeline to determine promoter-associated and differentially enriched CTCF motifs in CRC as compared to primary colonic tissue from ChIP-Seq datasets (**Chapter 3**).

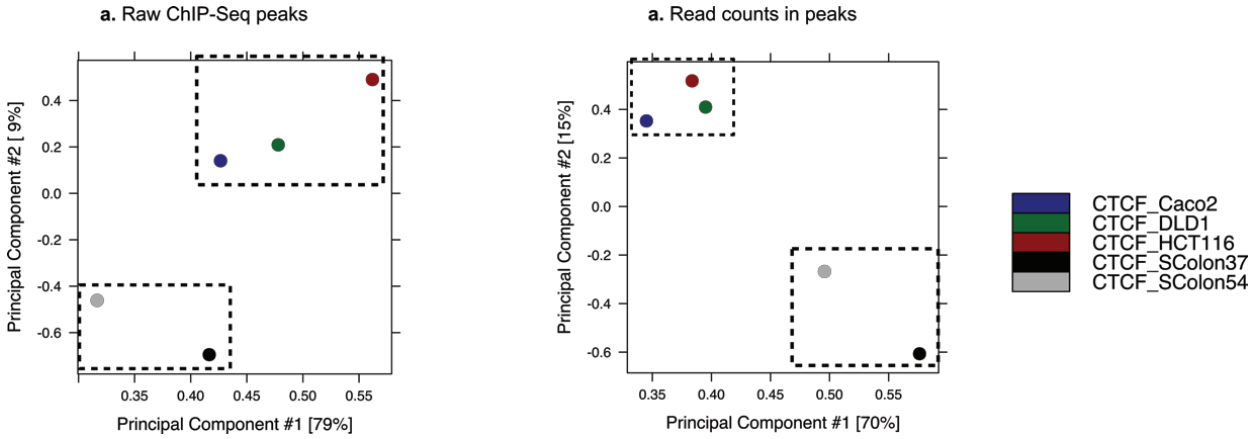


Figure 4-6: PCA correlation analysis on the CRC dataset. ***a****. Raw ChIP peaks, and* ***b****. Read count normalized peaks.*

PCA correlation analysis of peaksets was conducted on raw called peaksets and normalized to read counts prior to differential analysis, which improved the correlations between the replicates within the primary and CRC peaksets, respectively (**Figure 4-6**). Read count normalization datasets were used for subsequent analyses. Correlation analysis using Pearson correlation co-efficient analysis showed a clustering of normal and cancer peaksets before and after relative log expression normalization (**Figure 4-6a,b**). All the CTCF sites obtained from the CRC dataset were plotted on heatmaps to visualize the distribution of CTCF sites within 2 and 10 kb from hg38 annotated TSS' (**Figure 4.7**). Notably, the majority of CTCF sites within the CRC dataset are located proximal to TSS within the above-mentioned windows. Specifically, a majority of CTCF sites within these genomic ranges are located within genes (**Figure 4-7**) The distribution of these CTCF sites within these ranges is visually comparable to previous CTCF ChIP-Seq analysis studies revealing a majority CTCF sites located around transcriptional start sites [97].

Differential analysis of ChIP-Seq primary colon and CRC cell lines revealed 33 511 differentially enriched CTCF peaks in CRC with only 152 (0.5%) displaying significantly lower CTCF enrichment (LCe) in CRC (FDR<=0.05) (**Figure 4-8c**). Although, the most enriched motif (19.74%) in the LCe sites was the CTCF-like (CTCF-L/BORIS) motif (**Table 6-9**). Intriguingly, only 15 (7.89%) canonical CTCF motifs (MA1390.1) were found in 12 LCe sites (**Table 4-5**).

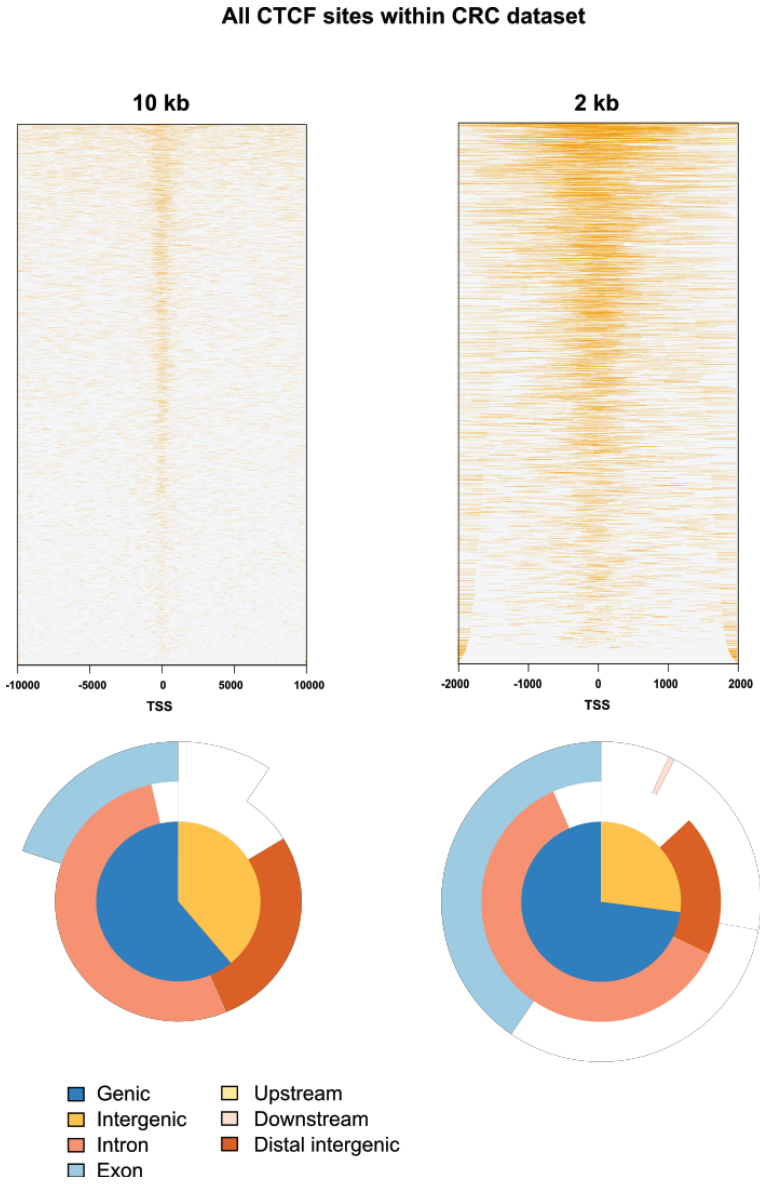


Figure 4-7: Heatmaps and vennpies of all CTCF sites in CRC dataset. Left panel*. within 10kb window from transcriptional start site (TSS) and a vennpie depicting the location of CTCF sites within this 10 kb window.* ***Right panel*** *within 2 kb window around TSS and vennipie describing the locations of CTCF sites within the 2kb window around TSSs. All annotations were annotated using the hg38 human genome in ChIPpeakAnno.*

The binding of CTCF to gene promoters validated in several examples, including at the promoter of tumor suppressor genes. Adequate transcriptional activity at these loci requires specific CTCF enrichment (**Section 4.4.1**). Genomic annotations of the LCe-CRC dataset described only 25 LCe sites located proximal (1 kb) to promoter-TSS's (**Figure 4-9**). Only 4 PA-LCe sites (5 motifs) were found to contain at least one canonical CTCF motif and were preferentially located proximal to (<10kb) antisense lncRNA loci (**Table 4-6**). Notably, the canonical CTCF site is considered indicative of CTCF binding in this ChIP-Seq dataset, however other variations in the genomic sequence of the CTCF motif (**Section 1.3.2**)[68]. These motif variations of the CTCF motif are outside the scope of this study.

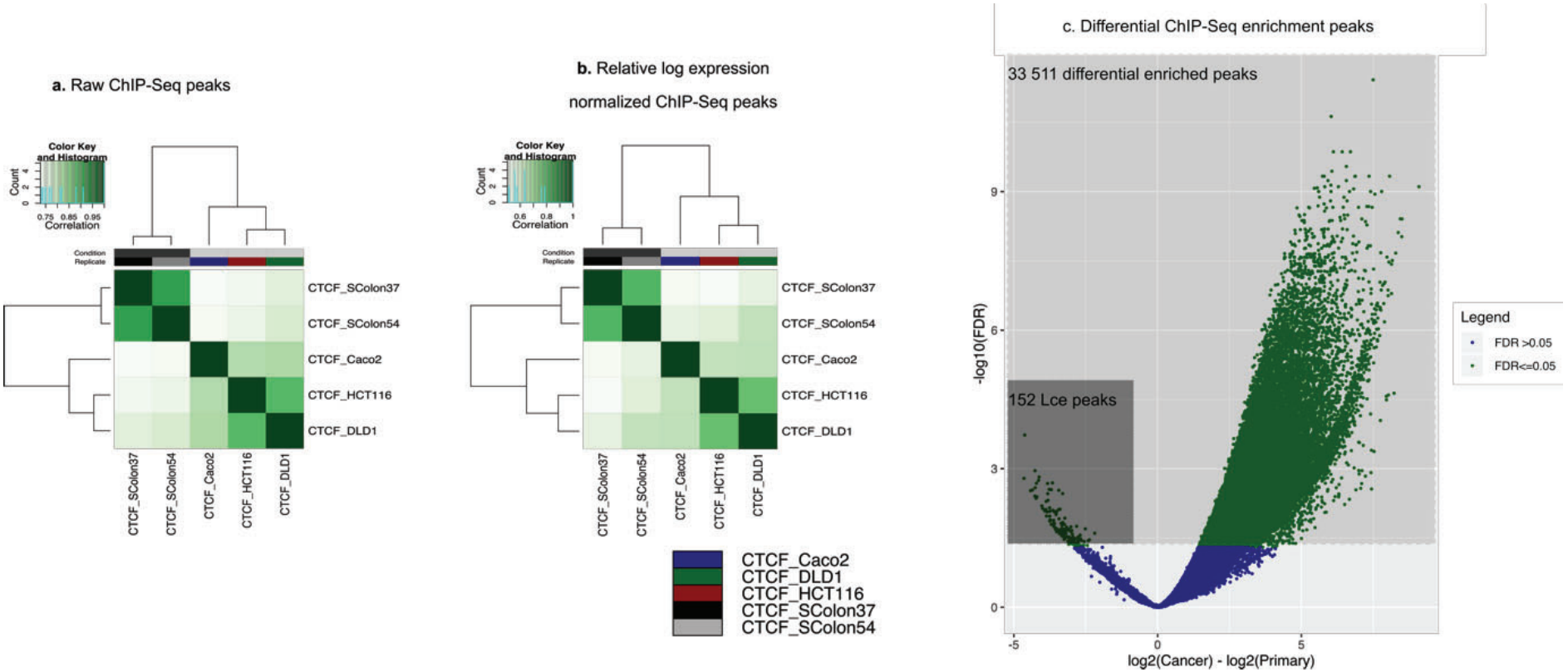


Figure 4-8: Differentially bound CTCF peaks in primary colon vs CRC cell lines.

Of the 3479 LCe peaks identified, 2648 (80%) were PA-LCes that, in general, were located at or proximal to bidirectional promoters expressing antisense lncRNAs with differential expression in CRC, including validated tumour-suppressive lncRNA, ZNF582-AS1. A pending question is whether these differences in CTCF binding are a consequence or a cause of the differential lncRNA expression observed in cancers.

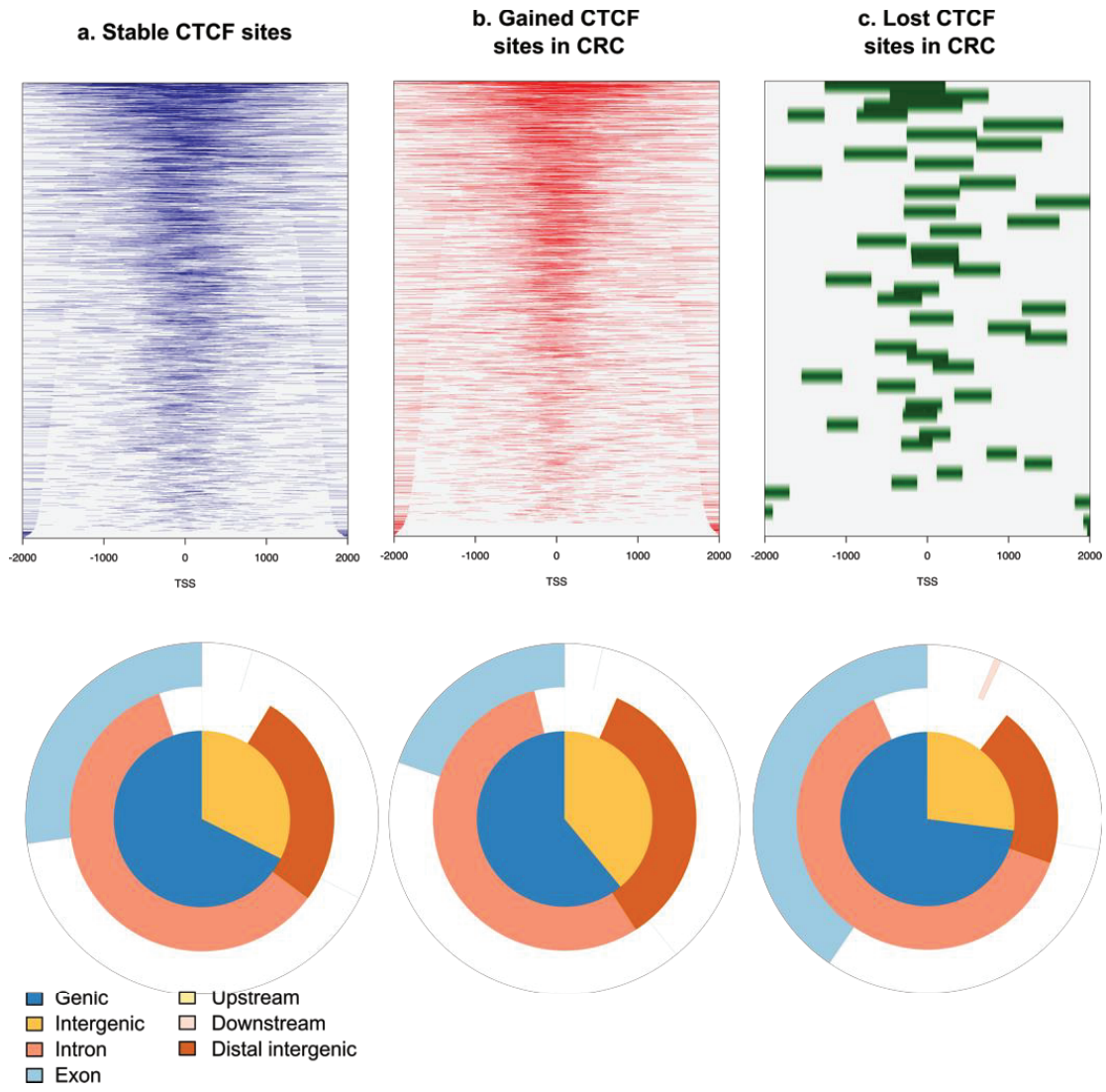


***Figure 4-9: Distribution of CTCF sites in CRC dataset**. **Top panel:** Heatmaps depict CTCF binding within 2kb of TSS as annotated by ChiPpeakAnno for **a.** stable, **b.** gained and **c.** lost CTCF sites in CRC as compared to primary sigmoid colon cells. **Bottom panel:** Vennpies annotating stable, lost and gained CTCF sites by their genomic location.*

CTCF binding at promoter regions has repeatedly been shown with 7-14% of silent promoter regions estimated to contain CTCF motifs[189] and over 50% of active promoter regions containing CTCF motifs in multiple cell types[97]. Therefore, it is not surprising that the annotation of all the CTCF sites in the CRC dataset were located within 2kb of the hg38 TSSs

as displayed in heatmaps annotated by ChIPpeakAnno (**Figure 4-7**). Vennpies of differentially analysed CTCF sites; stable, gained and lost in CRC demonstrate the preference of CTCF binding site location to genic regions proximal to TSSs as compared to intergenic regions (**Figure 4-9**). Full genomic location annotations of all CTCF sites in each of these datasets can be found in the bed files located within the Github project home page (PA-LCe-Discovery/CRC_Results/DiffPeaks/annotations).

Table 4-4: LCe sites in CRC peak annotations

Chr	Start	End	Strand	Annotation	Distance to TSS	Gene Name	Gene Type	CpG%	CD4+-*CTCF*-ChIP-Seq(Barski_et_al.)/ Homer Distance From Peak(sequence,strand,conservation)
chr19	56393007	56393624	+	promoter-TSS	286	ZNF582	protein-coding	0.074554	527(CAGCCGGAAGATGGCGCAGA,-,0.00)
chr17	29613033	29614064	+	exon	4569	CORO6	protein-coding	0.081474	207(CCTCCACTAGAGGGTGGTGG,-,0.00),546(GTAGCGCCCCCTGCCGGGAC,+,0.00)
chr11	63999714	64000939	+	exon	14473	OTUB1	protein-coding	0.093878	123(TGGGCAGAAGGGGGCAGGAA,-,0.00),863(AGGCCGGCAGGGGGCGCACC,-,0.00)
chr14	24310669	24311883	+	promoter-TSS	94	CIDEB	protein-coding	0.088962	1124(CACCCTCCAGAGGTCAGTGT,-,0.00)
chr12	124369295	124369776	+	intron	-32293	MIR6880	ncRNA	0.060291	281(TGGCCGGCAGAGGGAGGGGT,-,0.00)
chr17	7589144	7589922	+	exon	637	SOX15	protein-coding	0.077121	325(GCTGCGCCATCTGGCGGCGC,+,0.00)
chr5	123094964	123095784	+	intron	-5058	LOC105379152	ncRNA	0.080488	565(CTGCCAGCAGGCGGCACTCC,-,0.00)
chr10	35639704	35640204	+	exon	1298	MIR4683	ncRNA	0.094000	481(CACCCACCAGATGGAGCTGG,-,0.00)
chrX	138711902	138712456	+	promoter-TSS	72	FGF13-AS1	ncRNA	0.041516	356(ACGCTGCCACCCAGTGGTTT,+,0.00)
chr19	38869297	38870488	+	intron	8387	RINL	protein-coding	0.073048	700(GGAGCGCCACCTTGCGGCGC,+,0.00)
chr18	77899678	77900313	+	Intergenic	93732	LINC01029	ncRNA	0.059843	603(CGGCCGCTAGATGGCAGCGC,-,0.00)
chr2	185738760	185739273	+	promoter-TSS	77	LOC101927196	ncRNA	0.087719	256(GCGTCGCCCTCTGGCGGCCG,+,0.00),347(AGGCTGCCCTCTTGCGGCGC,+,0.00)

protein coding, miRNA -associated, ncRNA-associated, promoter-associated

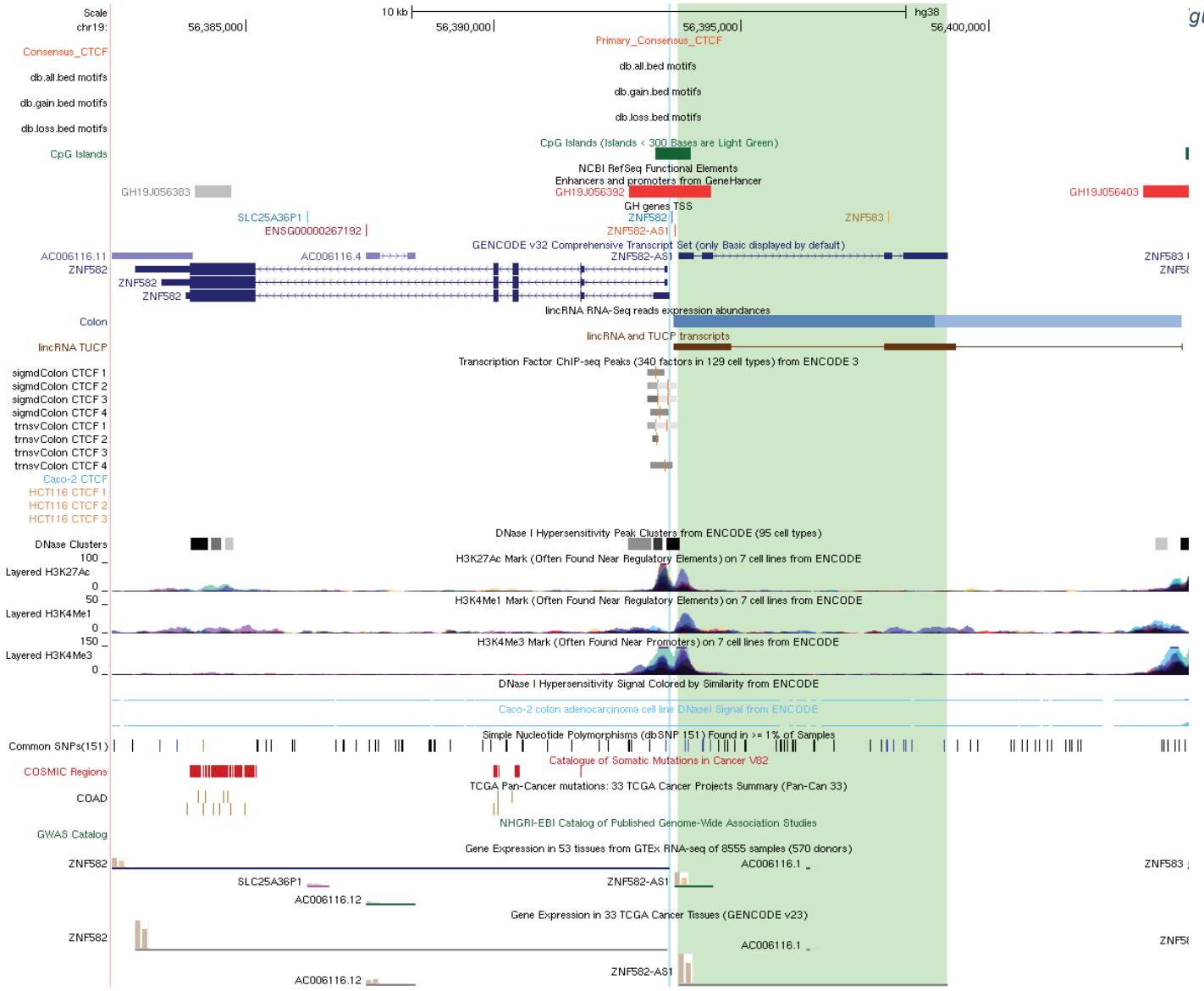
4.4.1 PA-LCe as-lncRNAs as potential tumor suppressors

4.4.1.1 PA-LCe at ZNF582-AS1 locus is a CRC meQTL

According to the UCSC Genome Browser (GRCh38), the PA-LCe site at the chr19:56393010-56393665 locus, is found at a bidirectional promoter which transcribes the *ZNF582* gene as well as an antisense transcript ZNF582-AS1 (**Figure 4-8**). According to RNA-Seq, the ZNF582-AS1 transcript is enriched in the colon as well as ubiquitously expressed in all tissues from the Human Body Map lincRNAs[190,191]. Both *ZNF582* and *ZNF582-AS1* are upregulated in COAD[192] (**Figure 4-8**). This PA-LCe CTCF motif, has CTCF ChIP-Seq peaks in both the transverse and sigmoid colon, which are lost in the Caco-2 and HCT116 cells (**Figure 4-8**). Furthermore, the PA-LCe CTCF motif is located within a CpG island, and DNAse I hypersensitivity peak cluster (95 cell types). This CTCF motif is located at a region displaying an active promoter histone signature in 7 ENCODE cell lines: high H3K27Ac, H3K4me3 and low H3K4me1. Intriguingly, this PA-LCe CTCF motif does not appear to contain COSMIC, TCGA or GWAS variants, however several variants have been annotated within the 2kb region surrounding this motif (**Figure 4-8**).

ZNF582-AS1 was identified as a target for epigenetic silencing in CRC using a lncRNA discovery pipeline that integrated H3K4me3 ChIP-Seq and reduced representation bisulphite sequencing (RRBS) data analysis[193]. In this study, the CpG islands within the *ZNF582-AS1* promoter and gene body were shown to be hypermethylated in CRC adenomas and cell lines as compared to normal colonic tissue, as well as others cancers[193]. RNA-Seq analysis revealed that although multiple ZNF582-AS1 variants were expressed in normal colonic tissue, the expression levels of all *ZNF582-AS1* variants was significantly downregulated in CRC[193].

Similar expression and methylation patterns were observed for *ZNF582* supporting previous data that identified the *ZNF582* gene as a predicator of prognosis in cervical[194], osophagael[195], renal[196] and colorectal cancer[193]. Furthermore, ectopic expression of *ZNF582-AS1* suppressed colony formation in CRC cell lines, RKO and SW480, but not HCT116 cells[193]. A recent differential expression analysis study also identified low expression of *ZNF582-AS1* as a prognostic marker in renal cancer [196]. Together these findings suggest that ZNF582-AS1 is a tumor suppressing lncRNA whose hypermethylation and transcriptional downregulation,



'gure 4-10: ZNF582 UCSC GRCh38 View

likely affecting CTCF enrichment, can be used as a diagnostic and prognostic marker for multiple cancers.

4.4.1.2 FGF13-AS1 is a tumor suppressor and prognostic marker

The PA-LCe CTCF motif located at the chrX:138711873-138712449 locus is located within intron 1 of *FGF13* and intron 1 of the antisense lncRNA *FGF13-AS1*(**Figure 4-9**). This CTCF motif is located within a DNAse I hypersensitivity cluster 25 bp downstream from a CpG island which is enriched for CTCF ChIP-Seq peaks in primary colonic tissue, which are lost in Caco-2 and HCT116 cell lines (**Figure 4-9**). Although proximal to an H3K27Ac peak in 7 ENCODE cell lines which aligns to the FGF13-AS1 locus, the motif itself is not enriched with H3K27Ac, H3K4me1 and H3K4me3 marks (**Figure 4-9**). Notably, the *FGF13/FGF13-AS1* PA-LCe CTCF motif follows the directionality of the *FGF13-AS1* gene (**Figure 4-9**).

The *FGF13* gene is upregulated in breast[197], cervical[198], prostate and colorectal cancer[199]. Notably, the *FGF13* gene was found to be hypomethylated in prostate cancer[200]. In CRC, post-transcriptional suppression of FGF13 by miR-10b, appears to suppress cell proliferation, migration and invasion in CRC cell lines[199]. FGF13 participates in a negative feedback loop with *TP53* and FGF13 resident miR-504[201]. Specifically, *p53* inhibits the transcription of FGF13 and miRNA 504[201], which emanates from FGF13 intron 2, while miR504 binds directly to the 3'UTR of *TP53* transcripts, inhibiting their expression in various cancers[202]. In CRC tumors, FGF13 expression is downregulated as compared to normal colonic tissue (**Figure 4-9**). Altogether this data suggests that FGF13 may function is a tumor suppressor.

Recently, FGF13-AS1 has also been implicated as tumor suppressor whose expression is negatively correlated with patient prognosis in breast cancer[203]. Specifically, the overexpression of FGF13-AS1 in MDBA-MB-231 cells suppressed cell proliferation, colony formation, cellular migration and MYC expression and stability. In MCF7 cells, FGF13-AS1 RNAi-mediated knockdown resulted in increased growth rates, invasion and migration, MYC expression and stability as well as xenograft tumor sizes and lung metastatic nodes in mice models. This function is mediated by FGF13-AS1's ability to repress glycolysis and the expression of stem cell markers *OCT4* and *SOX2* in MDBA-MB-231 cells[203]. Mechanistic assays including IGF2BP1 RNA immunoprecipitation (RIP), ChIP and RNAi suggest that FGF13-AS1 suppresses metastasis by functioning as a molecular sponge for IGF2 RNA

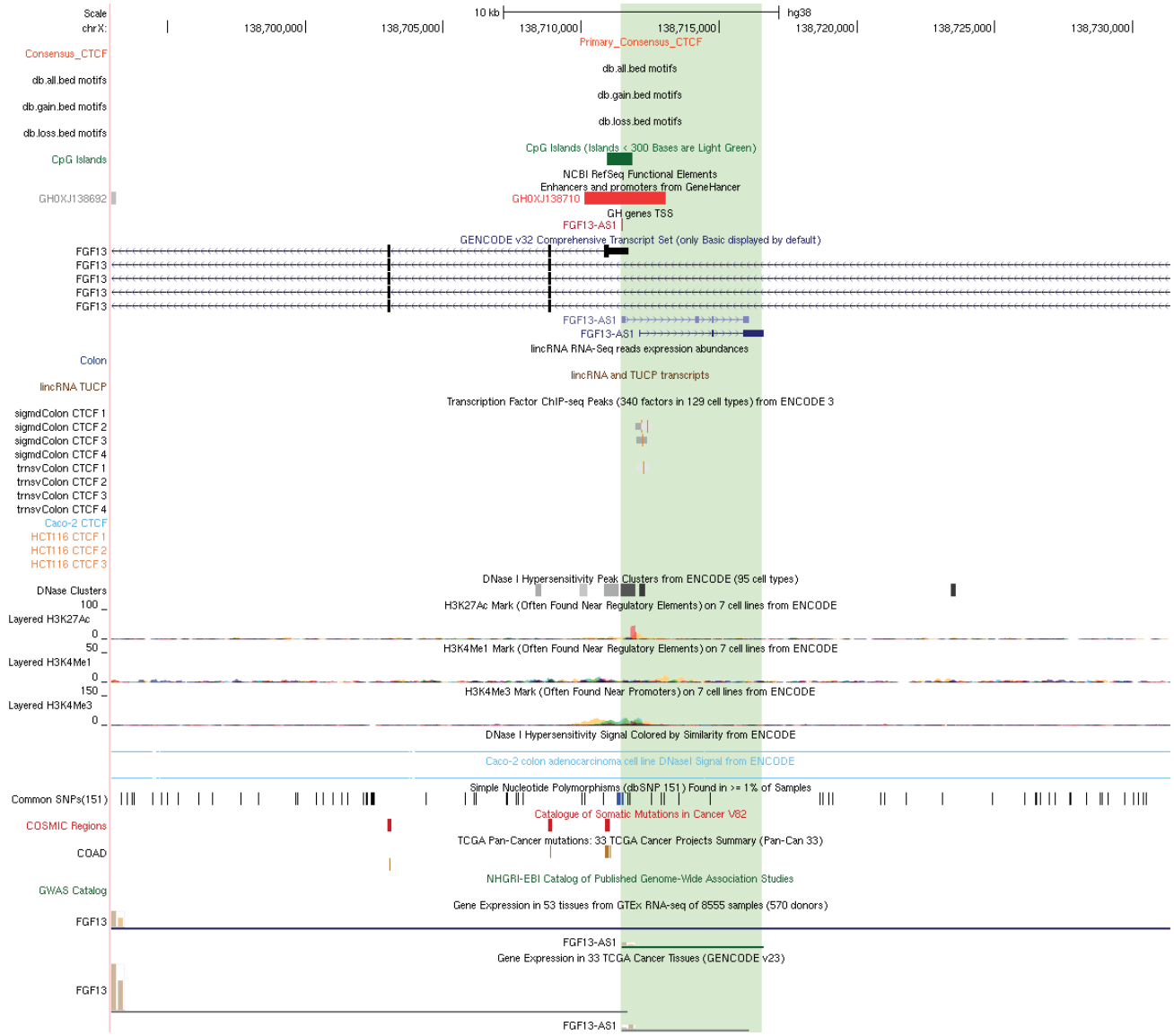


Figure 4-11: FGF13-AS1 UCSC Browser (GRCh38) Annotation.

binding proteins (IGF2BPs) preventing the interaction between MYC and IGF2BPs which in turn suppresses glycolysis and stemness in breast cancer[203]. Furthermore, Ma and colleagues postulated that FGF13-AS1 participates in a negative feedback loop with MYC expression where MYC binding on *FGF13-AS1* promoter prevents its transcription and binding to IGF2BP and thus promotes MYC-mediated oncogenesis[203]. Furthermore, PCAWG RNA-Seq data shows that FGF13-AS1 is expressed in normal colonic tissue but not in CRC tumors (**Figure 4.9**) and cell lines (**Figure 4-13**). Altogether this suggests FGF13-AS1 possesses tumor suppressive capabilities in various cancers.

The opposing transcriptional expression and functionality of FGF13 and its antisense transcript, FGF13-AS1, in cancer suggests that the inhibition or progression of oncogenesis is dependent on which of the divergent transcripts arising from this promoter are upregulated. Intriguingly, the *FGF13* gene body is hypomethylated in prostate cancer, providing a CTCF-binding permissive landscape while in this study CTCF binding at the FGF13 promoter is reduced. It is tempting to speculate that methylation-dependent the binding of CTCF at this locus may function as the "switch" that favours *FGF13*, over *FGF13-AS1*, expression in cancer.

4.4.1.3 FSIP2-AS2 (LOC101927196) as a potential tumor suppressor

The *FSIP2/FSIP2-AS2* locus contains two PA-LCe *CTCF* motifs separated by 91bp. These motifs emanate from intron 1 of FSIP2, intron 1 of *FSIP2-AS1* and within the *FSIP2-AS2* promoter region. Both the antisense lncRNA transcripts, FSIP2-AS1 and FSIP2-AS2 are transcribed in the 3'-5' direction while *FSIP2* is transcribed in the 5'-3 direction. Both *FSIP2/FSIP2-AS2* PA-LCe CTCF motifs share the directionality of the FSIP2 gene. The *FSIP2/FSIP2-AS2* PA-LCe CTCF motifs emanate from within a CpG and DNAse I hypersensitivity peak cluster (**Figure 4-10**). Similar to the FGF13/FGF13-AS1 PA-LCe motif, these motifs are enriched with CTCF in primary colonic cells but not CRC cell lines and do not appear to have the H3K27Ac, H3K4me1 and H3K4me3 markers in 7 ENCODE cell lines (**Figure 4-10**). Unlike the *FGF13/FGF13-AS1* PA-LCe motif, both the *FSIP2/FSIP2-AS2* PA-LCe *CTCF* motifs do not have COSMIC or TCGA mutations within >1kb of the *CTCF* motifs (**Figure 4-10**).

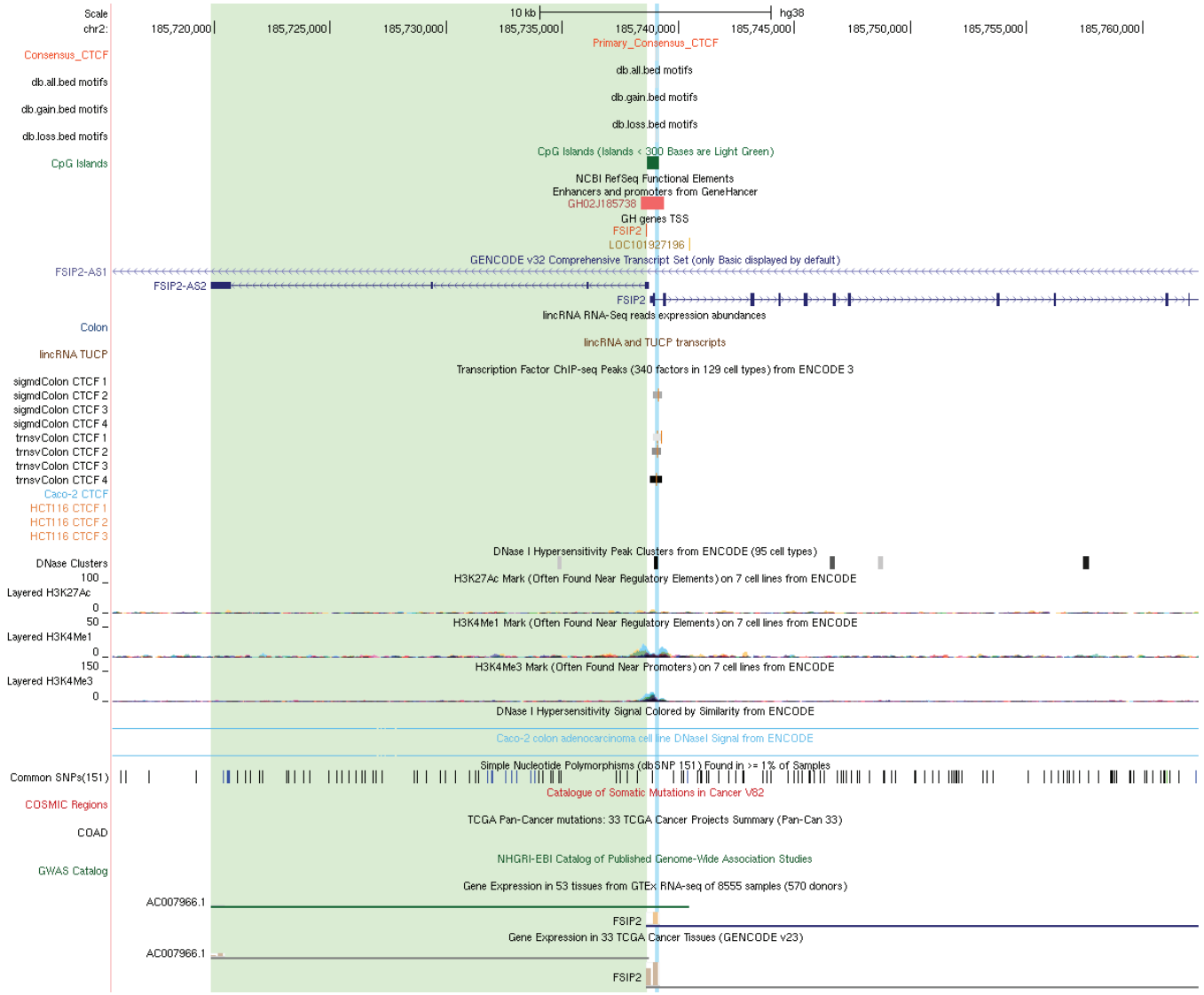


Figure 4-12: FSIP2/FSIP2-AS2 PA-LCe UCSC Genome Browser Annotations.

FSIP2 (fibrous shealth interacting protein 2) is a protein typically associated with the sperm fibrous-sheath that plays a role in spermatogenesis. The *FSIP2* gene is frequently mutated in metastatic breast carcinomas[204] and myeloid plasmacytomas[205] and has been presented as a "cancer driver" gene in breast intraepidermal adenocarcinomas[206]. Recurring amplifications at the FSIP2 gene have also been reported in testicular germ cell tumors[207].

To date, no studies have characterised FSIP-AS1 while the functionality of FSIP2-AS2 (LOC101927196) has only been described in rat autism models. In this model, FSIP2-AS2 was downregulated along with *Gsk-3β* and *Trx2* while *Fzd3*, *Wnt2*, *β-catenin*, *Bcl-2* and *Bax* were upregulated[208]. Notably these genes are involved in Wnt/β-catenin signalling and apoptotic pathways. Indeed, the FSIP2-AS2 overexpression in an autistic rat model lead to a decrease in cell proliferation and promotes apoptosis[208]. Thus, the authors postulate that FSIP2-AS2 attenuates Wnt signalling by upregulating FZD3, which leads to a decrease in β-catenin phosphorylation. Notably, the overexpression of FSIP2-AS2 suppresses the oxidative stress response in these cells, as measured by 4-hyrosxy-2-nonenal,(4-HNE), reactive oxygen species (ROS) and reactive nitrogen species (RNS) levels[208]. Although largely correlative, this data suggest that FSIP2-AS2 promotes apoptosis in a Wnt-signalling dependent manner. One can speculate that in the context of colorectal cancer, where *FZD3* is frequently upregulated[209,210], FSIP2-AS2 may function as a tumor suppressor.

4.4.1.4 PA-LCe CTCF motif at oncogenic and tumor suppressive CIDEB/LBT4R2 promoter

The chr14:24310670-24311881 is the locus that did not appear to contain an annotated lncRNA within 10kb of the PA-LCe CTCF motif. This PA-LCe CTCF site is located within bidirectional promoter that transcribes *LTB4R2* (Leukotriene B4 receptor 2) and *CIDEB* (Cell-death-inducing DFFA-like Effector B) in the 5'-3' and 3'-5' directions, respectively (**Figure 4-11**). Like, most PA-LCe CTCF motifs discovered in this study, this promoter is located within, or near a CpG island and DNAse I hypersensitivity cluster (**Figure 4-11**). This *CIDEB/LTB4R2* PA-LCe *CTCF* motif appears to be enriched with H3K27Ac and H3K4me1 7 ENCODE cell lines (**Figure 4-11**). Although no COSIMC or TCGA mutations align with the *CIDEB/LTB4R2*, several mutations are annotated in the 500bp region surrounding this PA-LCe *CTCF* motif (**Figure 4-11**).

LTB4R2 has been proposed as a prognostic marker in triple-negative breast cancer[211] and is frequently upregulated in oesophageal carcinomas where it contributes to malignant cell transformation[212]. Recently, LTB4R2 has been identified as a downstream target of the "cancer driver" KRAS pathway that mediates cell proliferation in *KRAS* mutant CRC cell lines LOVO and SW480[213]. Other studies have implicated LTB4R2 in mediating several aspects of cancer progression including cell proliferation, survival and metastasis in bladder, breast, pancreatic and prostate cancer[214–218]. Ultimately, these studies demonstrate the significant role *LTB4R2* plays in promoting oncogenesis.

Intriguingly, the *CIDEB* promoter is hypermethylated in lung, colon, endometrial and breast cancers[219–222] .The methylation status of the *CIDEB* promoter attenuates its transcriptional activity and has been correlated with poor patient prognosis in renal cell carcinomas[220,223]. It has been postulated that hypermethylation of the *CIDEB* promoter, and its downregulation, significantly contributes to the development of apoptosis resistance in cancer and is associated with poor patient prognosis [220,223]. Together, this data suggests CIDEB play a role in suppressing tumorigenesis.

Previous studies on this PA-LCe locus suggest that this bidirectional promoter transcribes both an oncogene (*LTB4R2*) and a tumor suppressor (*CIDEB*), and its expression is mediated by DNA methylation in a similar manner to the *TP53*/Wrap53 promoter. The opposing functionalities of the transcripts arising from this locus suggest that this PA-LCe, like the *FGF13*/FGF13-AS1 PA-LCe site, may function as a "cancer switch" where the progression or suppression of oncogenesis is dependent on which of the divergent transcripts emanating from the bidirectional promoter are upregulated. The mechanism that tilts the expression of this bidirectional promoter in either direction however, is currently unexplored.

4.4.2 PA-LCe as-lncRNA genomic features

Using multiple web-based genomic tools and browsers, we sought to characterize the defining features of the PA-LCe motifs discovered by our pipeline (**Table 4-6**). Like most CTCF motifs, all PA-LCe motifs were found within or proximal to; open chromatin (DNase I) hypersensitivity clusters, CpG islands (>300bp). However, PA-LCe CTCF motifs had varied histone marks in 7 ENCODE cell lines, which do not include cells of colonic origin (**Table 4-7**).

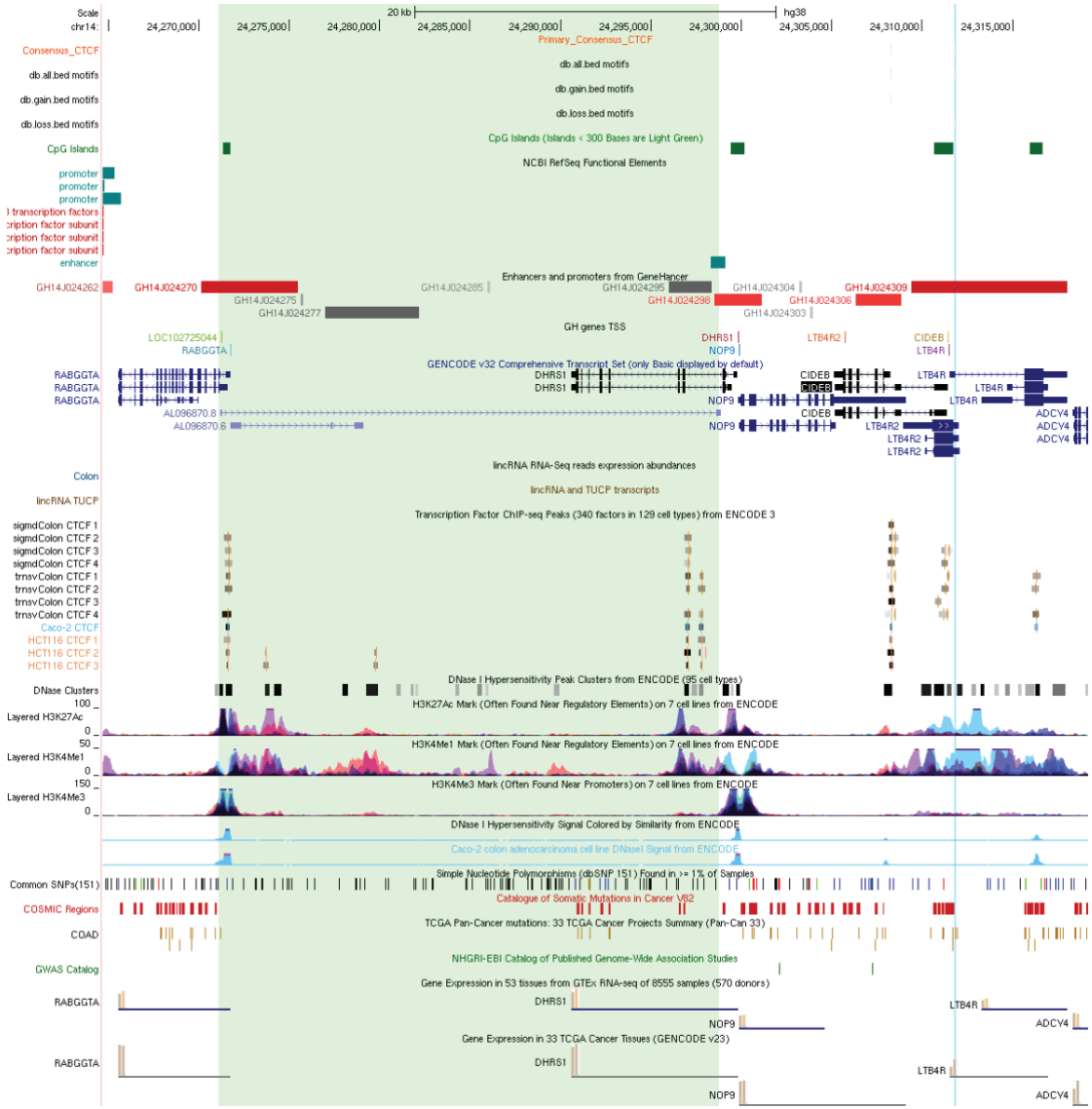


Figure 4-13: CIDEB/LTB4R2 UCSC Genome Browser Annotations.

Table 4-5: PA-LCe aslncRNAs in CRC

PA-LCe CTCF motif locus	PA-LCE aslncRNA	Strand	Genomic length	No. of exons	Transcript length
chr2:185738761 -185739281	FSIP2-AS2	-	18,848	4	1179
chr14:24310670 -24311881	CIDEB/ LTB4R2	-/ +	6236	7	2294
chr19:56393010 -56393665	ZNF582-AS1	+	5436	0	1606
chr: 138711873-138712449	FGF13-AS1	+	4,641 4,511	4 3	584 849

Table 4-6: Genomic Features of PA-LCe sites in CRC

PA-LCe CTCF motif locus	PA-LCE aslncRNA	PA-LCe CTC Motif ENCODE GRCh38 Epigenetic Features				
		DNAse I	CpG	H3K27Ac	H3K4me1	H3K4me3
chr2:185738761 -185739281	FSIP2-AS2	Internal	Internal	Low	Low	Low
chr14:24310670 -24311881	CIDEB/ LTB4R2	Internal	Proximal	Low	Low	Low
chr19:56393010 -56393665	ZNF582-AS1	Internal	Internal	High	Low	High
chard: 138711873-138712449	FGF13-AS1	Internal	Proximal	Low	Low	Low

4.4.3PA-LCe CTCF motifs in CRC are intra-TAD CTCF sites

The PA-LCe *CTCF* motif discovery pipeline developed in this study revealed only intra-TAD *CTCF* motifs as annotated by Hi-C maps of HCT116 cells containing an inducible RAD21-mAC vector prior to auxin induction[95] (**Figure 4-12**). Currently, no normal colon Hi-C data is available on the 3D genome browser, thus comparative chromatin contact visualization is not possible.

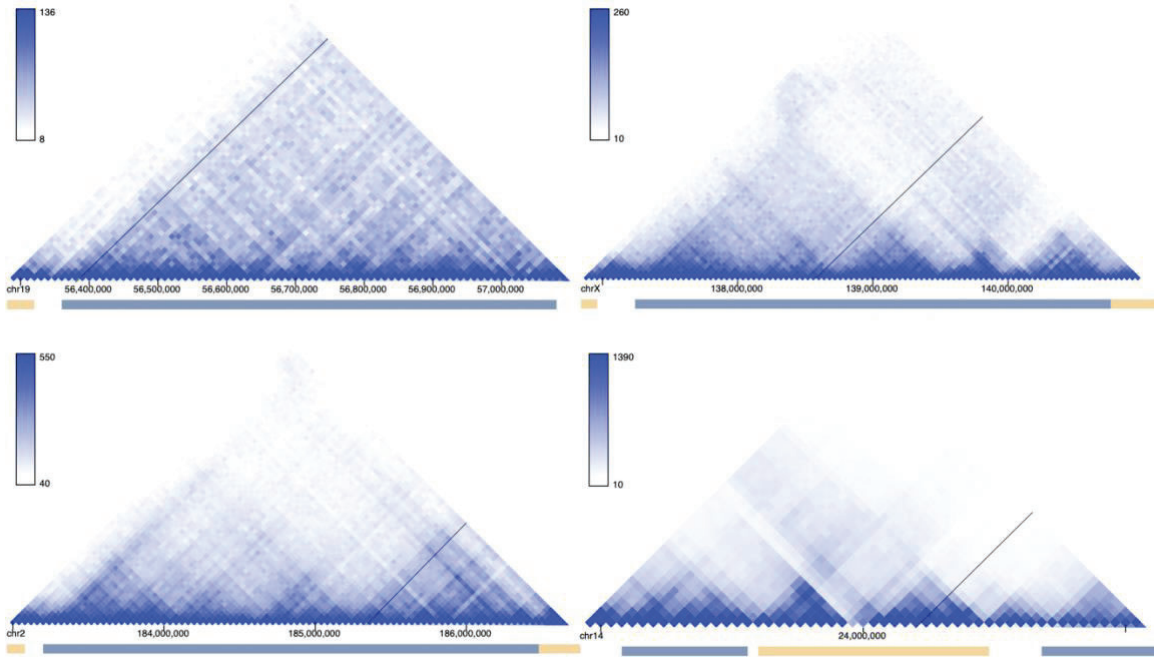


***Figure 4-14: Hi-C Maps of PA-LCe CTCF motifs in HCT116_RAD21-mAC_no_auxin at 40kb resolution[95,224]. a**. ZNF582, **b**. FGF13-AS1, **c**. FSIP2-AS2, d. LBT4R2. Black lines represent PA-LCe sites.*

4.4.4.PA-LCe PCAWG RNA-SEQ

The International Cancer Genome Project: Pan-Cancer Analysis of Whole Genomes RNA-Seq data demonstrates a general trend of transcriptional downregulation with both PA-LCe

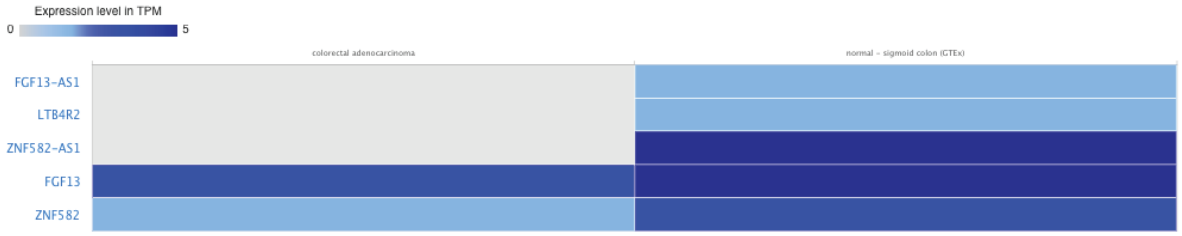


Figure 4-15: International Cancer Genome project: Pan-Cancer Analysis of Whole Genomes (PCAWG) RNA-Seq data of PA-LCe genes and asIncRNAs [192].proximal genes and asIncRNAs in CRC as compared to normal sigmoid colon (**Figure 4.13**). In general, PA-LCes proximal to asIncRNAs

are not expressed in CRC while genes proximal to PA-LCes are expressed in primary sigmoid colon, albeit downregulated. Notably, FSIP2, FSIP2-AS2 and CIDEB data is not available in this database.

4.5 Applying PA-LCe discovery pipeline using an ACC dataset

In order to determine the versatility of the PA-LCe pipeline developed in this study to identify tissue-specific PA-LCes in cancer, we applied a myelogenous leukaemia (ACC) dataset onto the pipeline. A pipeline validation dataset consisting of an leukaemia dataset underwent the PA-LCe discovery pipeline developed in this study. The ENCODE CTCF ChIP-Seq datasets included primary CD14+ monocytes and the B cell derived cell line GM12878 as primary control biological replicates. In this instance, the two ACC cell line K562 replicates were used as the cancer dataset (**Table 6-3**).

Dataset quality control results produced by FASTQC and ChIP-QC are shown in **Section 5.2**. Unlike the CRC dataset, the leukaemia dataset contained only SE read datasets. As with the CRC dataset, all steps of the PA-LCe pipeline with intermediary results are reported in **Section 5.2**. Similar to the CRC dataset PCA correlation analysis of peaksets was conducted on raw called peaksets and normalized to read counts prior to differential analysis, which improved the correlations between the replicates within the primary and ACC peaksets, respectively (**Figure 4-13**).

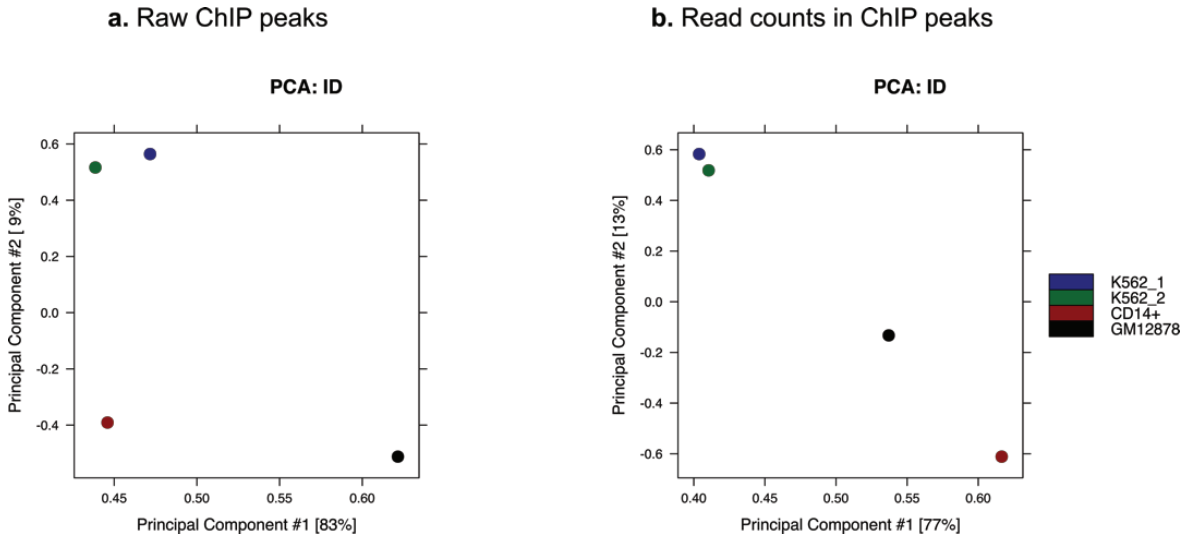


***Figure 4-16: PCA correlation analysis on the ACC dataset. a**. Raw ChIP peaks, and **b**. Read count normalized peaks.*

For the ACC dataset the PA-LCe pipeline revealed 11 717 differentially enriched *CTCF* peaks between primary and cancer datasets, with 3479 peaks displaying significantly lower *CTCF* enrichment (FDR<=0.5) (**Figure 4-13**). Genome-motif analysis did not identify the canonical CTCF motif in the as one of the Top 8 enrichment motifs (**Table 6-9**), however the canonical CTCF motif was discovered in 80% (2648/3479) of the LCe peaks. Of these, only 160 sites were annotated as PA-LCe sites while 31 PA-LCe sites were closest to annotated ncRNA TSSs, 9 of which were annotated closest to as-lncRNAs such as CASP8 And FADD Like Apoptosis Regulator (CFLAR) Melanotransferrin (MELTF) genes (**Section 5.2**).

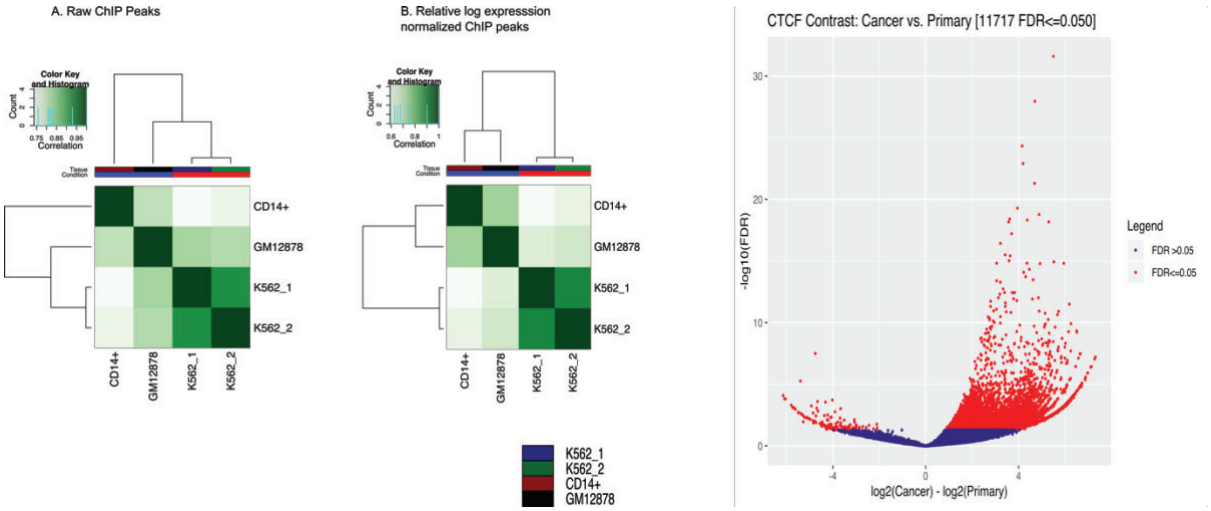


Figure 4-17: Differentially bound CTCF peaks in primary CD14+ monocytes and GM12878 vs K562 cell lines. ***a.*** *Raw ChIP-Seq Pearson correlation analysis of raw called peaksets.* ***b.*** *Relative log expression (DeSeq2) normalized called peaksets*

As with the PA-LCe sites discovered in the CRC dataset, the PA-LCe aslncRNAs in the ACC datasets shared similar genomic and epigenomic characteristics. Many of the as-lncRNAs discovered in this dataset have not been functionally characterized. Notably, some aslncRNAs like CFLAR-AS1 and Downregulated in Renal Carcinoma 3 (DIRC3) have been identified as prognostic markers for squamous cell[225] and renal carcinomas[226], respectively. Altogether, this data further reveals the applicability of the PA-LCe pipeline in identifying tissue-specific oncogenic aslncRNA loci with dysregulated CTCF binding in cancer which may be used as diagnostic and therapeutic markers.

Chapter 5 Conclusions and Perspectives

This study presents a comprehensive ChIP-Seq analysis pipeline to identify promoter-associated CTCF binding sites with differential, specifically abrogated, CTCF enrichment that may be hijacked by oncogenes in an attempt to favour cancer progression by modifying specific transcriptional programmes. Here we define protomer-associated lower CTCF-enrichment CBSs as canonical CTCF motifs within <1kb from an annotated protomer-TSS with significantly reduced CTCF enrichment in the cancer cell line datasets as compared to the wild type/primary dataset. In this workflow, we identify promoter-associated lower-CTCF enrichment sites in colorectal cancer cell lines as compared to primary colonic tissue from CTCF ChIP-Seq data. We find that highly significant lower-CTCF enrichment sites, containing the canonical MA0139.1 CTCF motif [75], are frequently proximal to bidirectional promoters of cancer-related genes and aslncRNAs.

Several tools and pipelines exist for ChIP-Seq analysis some of which are extensively described in **Chapter 3**. The alignment of ChIP-Seq reads to reference genomes is popularly conducted based on the Burrow-Wheeler transform employed by Bowtie2 and BWA. In this study, we utilized Bowtie2 for read alignments, which uses an inexact k-mer seeding strategy with a BLAST-like seed mapping approach (**Table 3-1**). ChIP-Seq peaks were then called with MACS2, a peak finding algorithm which uses a peak finding detection method that extends read tags in both the 5' and 3' direction before building a tag density landscape, using the maximum loci to predict DBP binding site locations (**Figure 3-3**). While several differential binding tools exist, in this study we followed differential binding analysis tool decision tree published by Steinhauser and colleagues [172] (**Figure 3-4**). In this study narrow CTCF peaks, with replicates from multiple ChIP-Seq experiments were classified as biological replicates with predefined region set. Thus, according to the Seinhauser *et al.* (2016) decision tree, we extensively reviewed ChIPComp, DBChIP, MMDiff and DiffBind tools applicable to the datasets used in this study (**Table 3-3**).

Here we opted to used DiffBind as a differential analysis tool and employed a DeSeq2 relative log expression normalization approach. This strategy allowed us to incorporate information from all MACS2 called peaks fitted into a negative binomial distribution using FDR<0.05 as a significance threshold between the primary and cancer datasets. Indeed, according to the

Seinhauser *et al.* (2016) decision tree, multiple tools could have been employed in our analysis, however, all the forementioned differential binding tools excluding DiffBind are applicable only to sharp/narrow peaks. As such to ensure maximum applicability of this pipeline to different dataset types i.e. sharp and broad peaks, we opted to use DiffBind. Furthermore, both DeSeq2 and edgeR normalization approaches can be used within DiffBind. This functionality allowed us to compare each of these normalization approaches with both the full and effective library sizes. Biologically, the number of differentially enriched CTCF sites between the primary and cancer conditions is expected to be high i.e. the number of stable/unchanged CTCF sites is low. This expectation was observed using the DeSeq2 normalization approach which as compared to edgeR, extends this assumption to subtract scaled control/background read counts from overlapping peaks (**Figure 4-5**). Thus, when using the full library size, which includes a greater proportion of these control reads by virtue of including all reads in bam files, DeSeq2 is able to more stringently identify high-confidence differential enrichment sites between conditions. As such the number of differentially enriched CTCF sites identified by the full library size are significantly lower than those identified using the effective library size (**Figure 4-5, Section 4.3.3**) fulfilling the biologically expectation. This strategy was effective in addressing research question posed in this study while attempting to control for the ChIP-Seq datasets obtained from varied experiments. The tools discussed here are by no means exhaustive and must be extensively reviewed in the context of the research question at hand.

Somatic mutations at CTCF sites can result in differential CTCF binding. However, redundancies at specific bases within the CTCF motif can mean some genomic mutations at CTCF motifs do not alter CTCF binding. Furthermore, varied combinations of mutations within the same motif can disproportionately affect CTCF binding at that genomic locus. Thus, while WGS approaches have been used to identify cancer driving CTCF motifs at base-pair resolution, these mutations cannot be fully prescriptive of differential CTCF binding and require the CTCF enrichment data availed by epigenetic-based analysis such as ChIP-Seq or ChIP-Exo. Previous strategies consisted of whole genome sequencing or association approaches to identify overrepresented mutational sequences by mutation calling followed by CTCF motif scanning at identified loci. Indeed these approaches have identified mutational hotspots in gastric[77] and colorectal cancers[78]. Additionally, while preparing this manuscript, a computational method CNCDriver was developed. This WGS-based approach identified CTCF mutational insulator driver sites in 1962 genomes in various cancers, including CRC[227]. The validation of CNCDriver discovered CTCF insulator sites, validated by CTCF ChIP-Seq,

3C and CRISPR-Cas9 and suggested that CTCF insulator sites function as putative oncogenic drivers [227].

GWAS approaches have also been previously employed identify CTCF-dependent CRC-specific driver mutations. For example the 1000G imputation and meta-analysis of 5 GWAS studies representing over 7000 CRC genomes identified 3 new susceptibility loci for CRC, with the second strongest association linked to the uncharacterised intergenic lncRNA, RP11-58A18.1[228] while a combinatorial GWAS, in situ promoter capture Hi-C (CHi-C), RNA-Seq and ChIP-Seq analysis approach in 34 627 CRC cases identified 31 new CRC risk loci primarily located at enhancer and promoter regions[229]. In this study using a DeSeq2 relative log expression normalization approach with the full read count library size of MACS2 called peaks, we identified 33 511 differentially enriched CTCF sites in CRC cell lines as compared to primary sigmoidal colon cells. Of these, we identified of 14 CRC-specific LCe canonical CTCF motifs and 4 PA-LCe sites. While this approach is evidenced by differential CTCF enrichment identified from ChIP-Seq data, the mutational status of these sites was not analysed and therefore cannot be assumed to be the underlying mechanism driving differential CTCF binding in these datasets as the susceptibility loci identified using WGS or GWAS approaches. Unlike ChIP-Seq datasets, WGS and GWAS datasets tend to have limited accessibility and typically rely on bespoke experimental datasets. Thus, to mitigate each of these limitations, our approach sought to use publicly available ChIP-Seq datasets to identify differentially enriched CTCF binding loci in CRC cells, that specifically aimed to identify reduced binding of CTCF at promoter-associated sites.

The comprehensive pipeline developed in this study takes as input ENCODE ChIP-Seq FASTQ datasets, discovers differentially enriched CTCF peaks with DeSeq2 in DiffBind and determines the promoter-associated CTCF motifs within these peaks. It is important to note that the DeSeq2 normalization strategy using the total library size does not attempt to control for technical bias in the pull-down efficiency of the ChIP-Seq experiments nor the likely varied antibodies used in each dataset. One caveat to our pipeline is that the collection ChIP-Seq datasets from various labs likely introduced technical variations compounded by the intrinsic variabilities in the ChIP-Seq protocol. This necessitates the requirement of robust normalization strategies. The integration of normalization methods to comparative ChIP-Seq analysis tools has become widely standardized, although several strategies exist. Here, we employ a relative log expression normalization strategy[168] with DeSeq2 implemented in

DiffBind[186] in pre-processed datasets. Multiple normalization strategies for RNA-Seq data have been developed to control of multiple dataset usage, however, the comparison of ChIP-Seq data is complicated by the different background noises, signal-to-noise ratios and antibody types in distinct experiments. Unfortunately, while extensive normalization strategies have been integrated into this pipeline, these biases cannot be fully mitigated when using datasets from different experiments. Several attempts to control for such biases have been conducted in the prior art including differential enrichment strategies employed by tools such as ChiPComp[173] however, this inherent caveat of ChIP-Seq analysis has not been fully mitigated to date. For the purposes of the research question addressed in this study, we sought to use DeSeq2's relative log expression normalization strategy[168] as it incorporates information from all peaks in each experiment to estimate a common dispersion parameter, is robust and allows for arbitrary mean-variance relationships allowing it to be highly adaptive to the different datasets (**Table 3-3**) used in this study.

In this pipeline, the ChIP-Seq peaks used for differential analysis are required to be called by MACS2 in more than one dataset i.e. only consensus peaks in the Sigmoid 37 and Sigmoid 54 dataset were used as control peaks while in the CRC cell lines, peaks were used for differential enrichment analysis only if said peak was identified in at least two datasets. This absolute requirement for multiple datasets per condition treated effectively as replicates for each condition in this pipeline may, albeit to a low extent, control for these artefacts. Thus, this pipeline can only be used for the differential analyses of ENCODE ChIP-Seq datasets with at least two biological replicates in two conditions (wild type and cancer) to discover known motifs with differential ChIP'ed DBP enrichment. This approach prevents, to a certain degree, false positives from being mis-interpreted as differentially enriched peaks, and further increases the stringency of this pipeline when identifying differentially enriched CTCF sites. Furthermore, the comparative sequencing depths of the datasets used in this study increased our confidence in the binding site discovery between these samples.

The PA-LCe discovery pipeline developed in this study resulted in some previously validated CTCF binding sites correlated with oncogenic activity. Intriguingly, the PA-LCe sites identified in this study emanate from bidirectional promoters whose differential methylation patterns in cancer have been previously described by others[193]**.** Genomic annotation analysis of the PA-LCe sites identified by this pipeline, revealed a subset of PA-LCes described as bidirectionally transcribed promoters whose antisense transcripts are lnRNAs. A survey of the literature

further implicates these aslncRNA PAL-Ces in oncogenesis (**Section 4.4**) . A notable example is the PA-LCe site identified in the promoter of tumour-suppressive aslncRNA ZNF582-AS1 [192]. The functionality of ZNF582-AS1 in CRC was previously discovered and described by bioinformatic pipeline that integrated RNA-Seq, ChIP-Seq and RRBS data in CRC tumors and cell lines in order to identify lncRNAs silenced by DNA methylation in CRC[193]. This data further strengthens the precision and robustness of which this PA-LCe discovery pipeline developed here. Coupled to these analyses, the discovery of cancer- and tissue-specific enhancer "docking-sites" promoter-proximal CTCF sites at oncogenes in various cancer cell lines by Richard Young's group, implicates PA-LCes as cancer-targets that elicit a therapeutic vulnerability[85]. Altogether, these studies further substantiate the relevance and usefulness of the PA-LCe CTCF discovery pipeline developed in this study in identifying potential oncogenic drivers.

The PA-LCe pipeline is applicable to multiple cellular contexts and can be used to identify candidate PA-LCe sites. In support of this, we used CTCF ChIP-Seq ACC cell lines and primary CD14+ monocytes and GM12878 to determine the applicability the PA-LCe pipeline developed in this study. Applying the developed PA-LCe pipeline to ChIP-Seq leukaemia datasets revealed similar genomic characteristics and antisense-lncRNAs to those identified in CRC. Together, this data reveals a potential mechanism in which CTCF enrichment at the promoters of cancer-targeted genes and/or aslncRNAs are targeted in cancer cells in order to promote oncogenesis.

The PA-LCe sites identified in this pipeline represent regions of differential epigenetic state, specifically CTCF binding. Whether the differential binding of CTCF at these loci is directly linked to the mutational status of the PA-LCe sties is yet to be determined. Future work will include the validation of these PA-LCe sites in matched normal versus tumor samples to determine the mutational status of PA-LCe sites in cancer using assays such as 3C, RRBS, Capture-C, CRISPR-Cas9 and the bacterial one-hybrid assay. Given the partially documented methylation status of PA-LCe sites from previous studies[193], it would be interesting to determine the effect of DNA methylation on these PA-LCe sites, which can be assayed using RRBS and/or reprogrammable CRISPR-mediated site-specific DNA methylation. Surveillance of RNA-Seq and CAGE datasets including FANTOM5, has shown that a majority of the PA-LCe proximal aslncRNAs are downregulated in CRC tumors. It would be interesting to

determine whether the expression of these aslncRNAs is directly linked to CTCF binding at their promoter regions within the same sample.

Recent studies implicating the formation of CTCF-lncRNA complexes at promoter regions as a mechanism of regulating CTCF docking on chromatin and subsequently promoter activity. From our data, this opens up the question on whether aslncRNAs at PA-LCe loci may be molecular targets for the regulation of CTCF binding and transcriptional activity in cancer. The discovery of aslncRNAs at the PA-LCe sites adds to the growing body of evidence that the non-coding genome regulates coding genome through multiple co-operative modes of action. It will be intriguing to determine whether the decrease in CTCF enrichment at PA-LCes is causal of the differential transcriptional activity of the promoters in which they reside.

Chapter 6 Appendices

Table 6-1: CRC Datasets

Dataset	SRA or ENCODE Accession ID(s)
Caco2_Control_1.1	SRR299299
Caco2_Test_3.1	SRR299297
DLD1_Control_1.1	DRR014664
DLD1_Test_3.1	DRR014660
HCT116_Control_1.1	SRR299382
HCT116_Test_3.1	SRR604580
SColon37_Control_1.1	ENCFF913JQM ENCFF024VTZ
SColon37_Test_1.1	ENCFF991UEV ENCFF547KHV
SColon54_Control_1.1	ENCFF848NLY ENCFF592JZV
SColon54_Test_1.1	ENCFF800GHL ENCFF100YUK

Table 6-2: Leukaemia dataset

Dataset	SRA, ENCODE or DDBJ Accession ID(s)
CD14_Control_1.1	SRR568246
CD14_Test_1.1	SRR568310
CD14_Control_1.2	SRR568247
CD14_Test_1.2	SRR568311
K562_Control_1.2	SRR5331211
K562_Test_1.1	SRR227523
K562_Control_1.3	SRR5331212
K562_Test_1.2	SRR227524

Table 6-3: List of tools used in PA-LCe discovery pipeline

GNU bash v5.0.2(1)
macs2 v2.1.2
bedtools v2.27.1
IGV v2.3.86
homer v3.4
fastq-dump v.2
fastq v0.11.3
bowtie2 v2.2.6
samtools v1.2
picard-2.21.3-0
R v3.5.1
pyranges v1.2.0
Rsamtools v1,34.1
rtracklayer v1.42.2
rgl v0.100.30
Tmisc v0.1.22
DiffBind v2.10.0
matrixStats v0.55.0
ReactomePA v1.26.0
VennDiagram v1.6.2.0
XVector v0.22.0
BiocParallel v2.10.0
AnnotationDbi v1.44.0

GenomeInfoDb v1.18.2

stats4 v3.5.1

stats v3.5.1

utils v3.5.1

bedr v1.0.7

edgeR v3.24.3

plot3Drgl v1.0.1

ChIPQC v1.18.2

ggplot2 v3.2.1

SummarizedExperiment v1.12.0

biomaRt v2.50.2

org.Hs.eg.db v3.70

futile.logger v1.4.3

Grid v3.5.1

TxDb.Hsapiens.UCSC.hg38.knownGene' v3.4.0

Biobase v2.42.0

IRanges v2.16.0

BiocGenerics v0.28.0

graphics v3.5.1

datasets v3.5.1

rstudioapi v0.10

limma v3.38.3

plot3D v1.3

RColorBrewer v1.1-2

DeSeq2 v1.22.2

DelayedArray v0.8.0

clusterProfiler v3.10.1

ChIPpeakAnno v3.16.1

ChIPSeeker v1.18.0

GenomicFeatures v1.34.8

GenomicRanges v1.34.0

S4Vectors v0.20.1

parallel v3.5.1

grDevices v3.5.1

methods v3.51

GRCh38 blacklisted regions
http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38/human/hg38.blacklist.bed.gz

Table 6-4: FASTQC Quality Metrics

Quality Metric	Threshold
Per base sequence quality	Lower quartile for any base must be more than 5
Per tile sequence quality	Any tile must show a mean Phred score less than 5, less than the mean for that base across all tile
Per sequence quality scores	The most frequently observed mean quality must be above 20 (1% error rate)
Per base sequence content	The difference between A and T, or G and C must be less than 20% in any position
Per base GC content	The sum of the deviations from the normal distribution must represent less than 30% of the reads.
Per base N content	Any position must show an N content of less than 20%
Sequence length distribution	Sequences must be the same length that is greater than 0bp
Sequence duplication levels	Non-unique sequences must make up less than 50% of the total.
Overrepresented sequences	Any sequence must represent less than 1% of total
Adaptor content	Any sequence must be present in less than 10% of all reads

Table 6-5 SAM File Format

Columns	Description
QNAME	Read name
FLAG	SAM flag
RNAME	Contig name or * for unmapped reads
POS	Mapped position of base q of a read on the reference sequence
MAPQ	Mapping quality
CIGAR	CIGAR string describing insertions and deletions
RNEXT	Name of mate
PNEXT	Position of mate
TLEN	Template length
SEQ	Read sequence
QUAL	Read quality
TAGS	Additional information in TAG

Table 6-6: BAM File Format

Columns	Description
Magic	BAM magic string
l_text	Length of header text
Text	Plain header text in SAM
N_ref	Number of reference sequences
L_name	Length of the reference name
Name	Refence sequence name
l_ref	Length of refence sequence
Block_size	Total length of alignment record, excluding this field
refID	Refence sequence ID, -1 for unmapped reads
pos	Length of read name POS
L_read_name	Length of read name QNAME
Mapq	Mapping quality MAPQ
Bin	BAI index bin
N_cigar_op	Number of operations in CIGAR
Flag	Bitwise flags FLAG
L_seq	Length of SEQ
Next_refID	Ref-ID for next segment
Next_pos	0-based leftmost pos of the next segment
Tlen	Template length TLEN
Read_name	Read name
Cigar	CIGA string
Seq	Read sequence

Qual	Phred-scaled base qualities
Tag	Two-character tag
Val_type	Value type
value	Tag value

Table 6-7: Browser Extensible Data (BED) Format

Field	Field requirement	Description
chrom	Required	name of the chromosome or scaffold. Any valid seq_region_name can be used, and chromosome names can be given with or without the 'chr' prefix
chromStart		start position of feature in standard chromosomal coordinates (i.e. first base is 0)
chromEND		position of feature in standard chromosomal coordinates
name	Optional	sequence label
score		a score between 0-1000
strand		defined as + (forward) and – (reverse)
thickStart		coordinate at which to start drawing the feature as a solid rectangle
ThickEnd		coordinate at which to stop drawing the feature as a solid rectangle
itemRgb		an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value of itemRgb set to "on"
blockCount		the number of sub-elements (e.g. exons or SNPS) within the feature
blockSizes		the size of sub-elements displayed in blockCount
blockStarts		the start co-ordinate of sub-elements

Table 6-8: General Feature Format (GFF) or General Transfer Format (GTF)

Field	Description
seqname	name of the chromosome or scaffold (Ensembl identifier); chromosome names can be given with or without the 'chr' prefix
source	name of the program that generated this feature, or the data source
feature	feature type name, e.g. Gene, Variation, Similarity
start	start position of the feature, with sequence numbering starting at 1
end	end position of the feature, with sequence numbering starting at 1
score	a floating point value
	defined as + (forward) or - (reverse)
frame	one of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on.
attribute	a semicolon-separated list of tag-value pairs, providing additional information about each feature

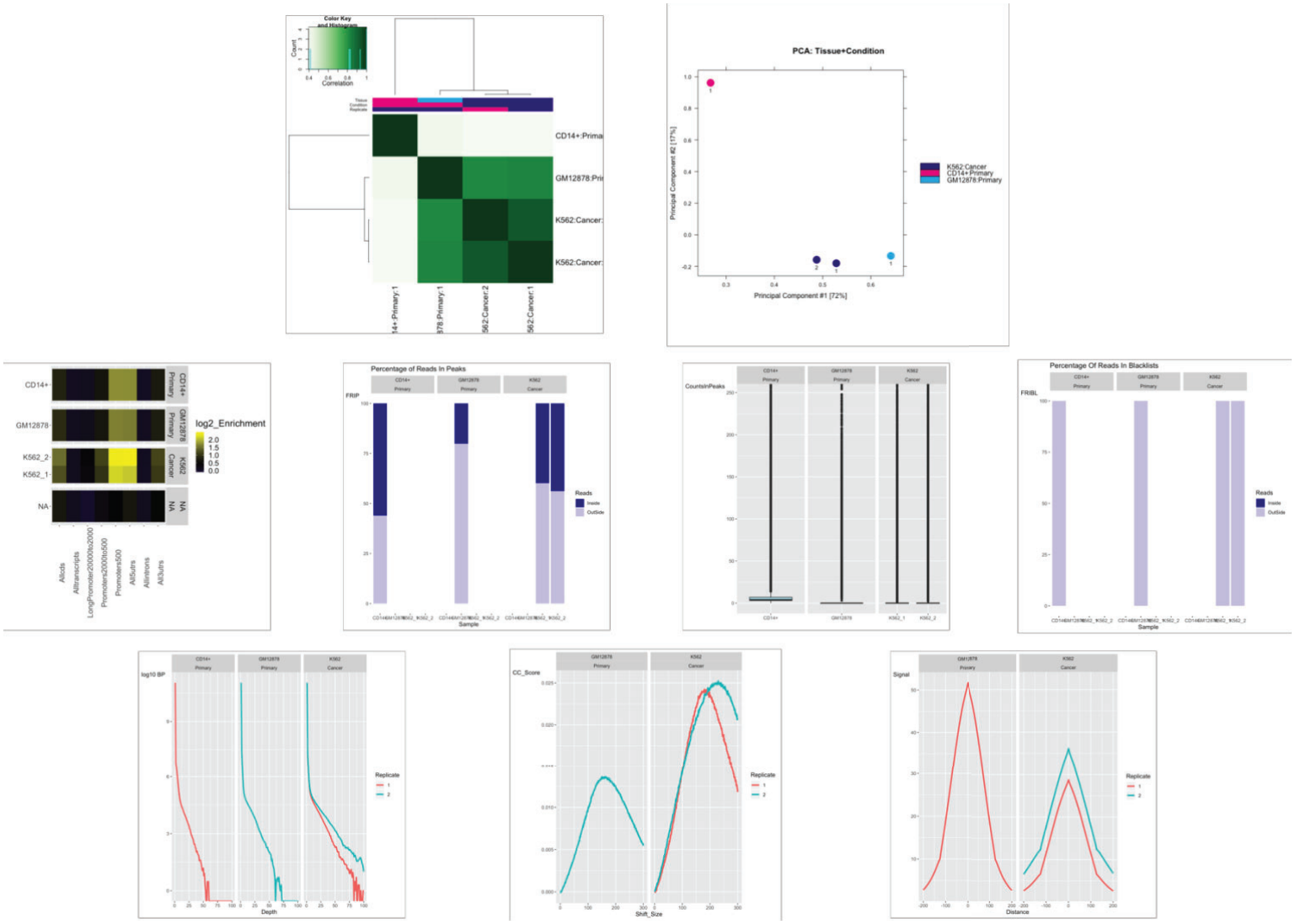


Figure 0-1: ChIPQC Results for leukaemia dataset.

profile across peaks.

Table 0-9: Top enriched motifs in CRC dataset; All peaks; Increased CTCF enrichment peaks; Lower CTCF enrichment peaks; Non-differentially bound peaks

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-1.300e-05	0.12%	1.34%	162.5bp (361366.9bp)	CTCFL/MA1102.1/Jaspar(0.857) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	-0.000e+00	2.11%	20.75%	150.5bp (3624189.4bp)	CTCF/MA0139.1/Jaspar(0.832) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-0.000e+00	20.14%	93.44%	284.5bp (3965611.6bp)	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer(0.928) More Information | Similar Motifs Found	motif file (matrix)

Top enriched motifs in CRC dataset; All peaks; Increased CTCF enrichment peaks; Lower CTCF enrichment peaks; Non-differentially bound peaks

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-0.000e+00	6.63%	41.94%	145.7bp (3859901.1bp)	CTCF/MA0139.1/Jaspar(0.830) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	-0.000e+00	6.60%	72.97%	209.5bp (3739457.2bp)	CTCFL/MA1102.1/Jaspar(0.859) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-0.000e+00	17.19%	92.70%	259.2bp (3591556.1bp)	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer(0.863) More Information | Similar Motifs Found	motif file (matrix)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-1.116e+00	1.96%	1.27%	103.0bp (1656235.1bp)	POL001.1_MTE/Jaspar(0.687) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	-5.764e-01	0.65%	0.50%	0.0bp (695068.5bp)	ZBTB33(Zf)/GM12878-ZBTB33-ChIP-Seq(GSE32465)/Homer(0.623) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-2.120e-01	0.65%	1.04%	0.0bp (1480624.3bp)	ERF069/MA0997.1/Jaspar(0.785) More Information | Similar Motifs Found	motif file (matrix)
4 *		1e0	-1.666e-02	1.31%	4.65%	15.5bp (1990163.4bp)	CTCFL/MA1102.1/Jaspar(0.747) More Information | Similar Motifs Found	motif file (matrix)
5 *		1e0	-5.407e-03	1.31%	5.39%	0.0bp (3385128.8bp)	PB0199.1_Zfp161_2/Jaspar(0.726) More Information | Similar Motifs Found	motif file (matrix)
6 *		1e0	-6.120e-04	5.23%	14.35%	78.2bp (1617098.5bp)	FRS9(ND)/col-FRS9-DAP-Seq(GSE60143)/Homer(0.752) More Information | Similar Motifs Found	motif file (matrix)
7 *		1e0	-0.000e+00	4.58%	85.06%	210.9bp (4840465.5bp)	SeqBias: CA-repeat(0.894) More Information | Similar Motifs Found	motif file (matrix)
8 *		1e0	0.000e+00	7.84%	56.39%	387.6bp (2897346.6bp)	SUT1(MacIsaac)/Yeast(0.774) More Information | Similar Motifs Found	motif file (matrix)
9 *		1e0	0.000e+00	0.00%	51.29%	0.0bp (4476045.7bp)	LARK(RRM,Znf)/Drosophila_melanogaster-RNCMPT00097-PBM/HughesRNA(0.768) More Information | Similar Motifs Found	motif file (matrix)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-0.000e+00	5.55%	62.84%	221.7bp (3451690.1bp)	SeqBias: CG bias(0.905) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	0.000e+00	8.90%	42.48%	166.4bp (4084456.8bp)	CTCF/MA0139.1/Jaspar(0.806) More Information | Similar Motifs Found	motif file (matrix)

Table 0-10: Top enriched motifs in leukaemia dataset; All peaks; Increased CTCF enrichment peaks; Lower CTCF enrichment peaks; Non-differentially bound peaks

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-0.000e+00	6.63%	41.94%	145.7bp (3859901.1bp)	CTCF/MA0139.1/Jaspar(0.830) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	-0.000e+00	6.60%	72.97%	209.5bp (3739457.2bp)	CTCFL/MA1102.1/Jaspar(0.859) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-0.000e+00	17.19%	92.70%	259.2bp (3591556.1bp)	Maz(Zf)/HepG2-Maz-ChIP-Seq(GSE31477)/Homer(0.863) More Information | Similar Motifs Found	motif file (matrix)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-3.590e-04	1.16%	3.55%	206.7bp (996951.9bp)	BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer(0.915) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	-0.000e+00	15.93%	93.80%	311.4bp (3442857.9bp)	KLF15/MA1513.1/Jaspar(0.802) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-0.000e+00	6.35%	93.64%	311.2bp (3530404.3bp)	P0510F09.23/MA1030.1/Jaspar(0.797) More Information | Similar Motifs Found	motif file (matrix)
4 *		1e0	0.000e+00	0.62%	39.21%	271.1bp (3329847.7bp)	DPL-1(E2F)/cElegans-Adult-ChIP-Seq(modEncode)/Homer(0.891) More Information | Similar Motifs Found	motif file (matrix)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e-3	-7.959e+00	6.25%	0.88%	109.0bp (2046561.4bp)	HIC1(Zf)/Treg-ZBTB29-ChIP-Seq(GSE99889)/Homer(0.708) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e-2	-5.052e+00	4.38%	0.60%	169.5bp (1469415.0bp)	U2AF2(RRM)/Homo_sapiens-RNCMPT00079-PBM/HughesRNA(0.728) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-8.724e-01	3.75%	3.04%	144.2bp (2637152.9bp)	UGA3/MA0410.1/Jaspar(0.769) More Information | Similar Motifs Found	motif file (matrix)
4 *		1e0	-3.293e-01	1.25%	1.68%	41.0bp (1772998.3bp)	CTCFL/MA1102.2/Jaspar(0.727) More Information | Similar Motifs Found	motif file (matrix)
5 *		1e0	-3.293e-01	1.25%	1.70%	146.0bp (1142452.7bp)	BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer(0.700) More Information | Similar Motifs Found	motif file (matrix)
6 *		1e0	-5.376e-03	1.25%	5.13%	106.5bp (2154846.4bp)	RDS1/MA0361.1/Jaspar(0.819) More Information | Similar Motifs Found	motif file (matrix)
7 *		1e0	-0.000e+00	1.25%	18.39%	97.0bp (3548257.0bp)	SKN7/SKN7_H2O2Lo/[](Harbison)/Yeast(0.755) More Information | Similar Motifs Found	motif file (matrix)
8 *		1e0	-0.000e+00	25.62%	65.71%	126.5bp (3976305.3bp)	FUS(RRM)/Homo_sapiens-RNCMPT00018-PBM/HughesRNA(0.958) More Information | Similar Motifs Found	motif file (matrix)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1 *		1e0	-3.100e-05	0.08%	1.20%	261.9bp (364209.3bp)	CTCFL/MA1102.2/Jaspar(0.891) More Information | Similar Motifs Found	motif file (matrix)
2 *		1e0	-0.000e+00	0.80%	37.52%	221.7bp (3117830.0bp)	RAP2-6/MA1052.1/Jaspar(0.895) More Information | Similar Motifs Found	motif file (matrix)
3 *		1e0	-0.000e+00	8.94%	50.70%	141.6bp (3791064.4bp)	CTCF/MA0139.1/Jaspar(0.808) More Information | Similar Motifs Found	motif file (matrix)

Chapter 7 Ethics approval and consent to participate

7.1 Ethics approval and consent to participate

Not applicable.

7.2 Availability of data and material

Project name: **PA-LCe-Discovery**

Project home page: https://github.com/LorettaM/PA-LCe-Discovery

Operating system(s): Mac

Programming languages: Linux and R

Datasets: Available from the NCBI: https://www.ncbi.nlm.nih.gov.

7.3 Competing interests

The authors declare that they have no competing interests.

7.4 Funding

This work was funded by the National Research Foundation of South Africa, grant number 86934.

Chapter 8 References

1. Misteli, T. Beyond the Sequence: Cellular Organization of Genome Function. Cell 128, 787–800 (2007).

2. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature Reviews Genetics 2, 35066075 (2001).

3. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, N.Y.) 326, 289–93 (2009).

4. Bolzer, A. et al. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. Plos Biol 3, e157 (2005).

5. Lieberman-Aiden, E. et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009).

6. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. Elife 6, e25776 (2017).

7. Nora, E. P. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–5 (2012).

8. de Wit, E. et al. CTCF Binding Polarity Determines Chromatin Looping. Mol Cell 60, 676–684 (2015).

9. Rao, S. S. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).

10. Ong, C.-T. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. Nature reviews. Genetics 15, 234–46 (2014).

11. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Gene Dev 20, 2349–2354 (2006).

12. Wendt, K. S. et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451, 796 (2008).

13. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc National Acad Sci 112, E6456–E6465 (2015).

14. Seitan, V. C. et al. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Research 23, 2066–2077 (2013).

15. Lupiáñez, D. G. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012–1025 (2015).

16. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 529, 110 (2016).

17. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 351, 1454–1458 (2016).

18. Rocha, S. T. da & Heard, E. Novel players in X inactivation: insights into Xist-mediated gene silencing and chromosome conformation. Nat Struct Mol Biol 24, 197–204 (2017).

19. Iyer, K. V. et al. Modeling and Experimental Methods to Probe the Link between Global Transcription and Spatial Organization of Chromosomes. PLoS ONE 7, e46628 (2012).

20. Boehning, M. et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. Nat Struct Mol Biol 25, 833–840 (2018).

21. Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M. S. & Mhlanga, M. M. Chromosomal Contact Permits Transcription between Coregulated Genes. Cell 155, 606–20 (2013).

22. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12, 7–18 (2010).

23. Ungerbäck, J. et al. Pioneering, chromatin remodeling, and epigenetic constraint in early T-cell gene regulation by SPI1 (PU.1). Genome Res 28, 1508–1519 (2018).

24. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. Trends Genet 27, 465–474 (2011).

25. Lambert, S. A. et al. The Human Transcription Factors. Cell 172, 650–665 (2018).

26. Bird, A. DNA methylation patterns and epigenetic memory. Genes & Development 16, 6–21 (2002).

27. Jones, P. A. & Takai, D. The Role of DNA Methylation in Mammalian Epigenetics. Science 293, 1068–1070 (2001).

28. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet 18, 517–534 (2017).

29. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci 31, 89–97 (2006).

30. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, D1005–D1012 (2018).

31. L., M. LncRNA discovery in the Listeria monocytogenes infection model. (2015).

32. Wang, K. C. & Chang, H. Y. Molecular Mechanisms of Long Noncoding RNAs. Molecular Cell 43, 904–914 (2011).

33. Magagula, L., Gagliardi, M., Naidoo, J. & Mhlanga, M. Lnc-ing inflammation to disease. Biochemical Society Transactions 45, 953–962 (2017).

34. Barichievy, S., Magagula, L., Shibayama, Y. & Mhlanga, M. M. Non-coding RNAs and Inter-kingdom Communication. 27–52 (2016) doi:10.1007/978-3-319-39496-1_2.

35. Lai, F. et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature 494, 497–501 (2013).

36. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299–308 (1981).

37. Snetkova, V. & Skok, J. A. Enhancer talk. Epigenomics-uk 10, 483–498 (2018).

38. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet 20, 437–455 (2019).

39. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167, 1170–1187 (2016).

40. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biology 16, 144–54 (2015).

41. Williamson, I., Lettice, L. A., Hill, R. E. & Bickmore, W. A. Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity. Development 143, 2994–3001 (2016).

42. Benabdallah, N. S. et al. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. Molecular Cell (2019) doi:10.1016/j.molcel.2019.07.038.

43. Deng, W. et al. Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. Cell 149, 1233–1244 (2012).

44. Bartman, C. R., Hsu, S. C., Hsiung, C. C., Raj, A. & Blobel, G. A. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. Mol Cell 62, 237–247 (2016).

45. Kim, T.-K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. Cell 162, 948–59 (2015).

46. Hughes, J. R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet 46, 205–212 (2014).

47. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database : the journal of biological databases and curation 2017, (2017).

48. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research 46, D794–D801 (2018).

49. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic Acids Res 44, D164–D171 (2016).

50. Lizio, M. et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology 16, 22 (2015).

51. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology 28, 817–25 (2010).

52. Ørom, U. et al. Long Noncoding RNAs with Enhancer-like Function in Human Cells. Cell 143, 46–58 (2010).

53. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182–7 (2010).

54. Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P. & Hayashizaki, Y. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. Sci Data 4, 170113 (2017).

55. Consortium, T. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).

56. Scruggs, B. S. et al. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol Cell 58, 1101–12 (2015).

57. Dao, L. T. et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nature Genetics 49, 1073–1081 (2017).

58. Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Reports 15, 2038–49 (2016).

59. Raj, A., Bogaard, P. van den, Rifkin, S. A., Oudenaarden, A. van & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods 5, 877–9 (2008).

60. Beliveau, B. J. et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. P Natl Acad Sci Usa 109, 21301–6 (2012).

61. Papantonis, A. et al. Active RNA polymerases: mobile or immobile molecular machines? Plos Biol 8, e1000419 (2010).

62. Morrison, J. A., McKinney, M. & Kulesa, P. M. Resolving in vivo gene expression during collective cell migration using an integrated RNAscope, immunohistochemistry and tissue clearing method. Mech Develop 148, 100–106 (2017).

63. Schoenfelder, S. et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet 42, 53–61 (2010).

64. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. Science 295, 1306–1311 (2002).

65. Hashimoto, H. et al. Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. Mol Cell 66, 711-720.e3 (2017).

66. Eom, K. S., Cheong, J. S. & Lee, S. J. Structural Analyses of Zinc Finger Domains for Specific Interactions with DNA. J Microbiol Biotechn 26, 2019–2029 (2016).

67. Yin, M. et al. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. Cell Res 27, 1365–1377 (2017).

68. Nishana, M. et al. Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation. Genome Biol 21, 108 (2020).

69. Saldaña-Meyer, R. et al. RNA Interactions Are Essential for CTCF-Mediated Genome Organization. Molecular Cell (2019) doi:10.1016/j.molcel.2019.08.015.

70. Hansen, A. S. et al. Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. Molecular Cell (2019) doi:10.1016/j.molcel.2019.07.039.

71. Network, T. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330 (2012).

72. Lobanenkov, V. V. et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene 5, 1743–53 (1990).

73. Klenova, E. et al. CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Mol Cell Biol 13, 7612–7624 (1993).

74. Kim, T. H. et al. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell 128, 1231–1245 (2007).

75. Barski, A. et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129, 823–837 (2007).

76. Li, Y. et al. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. Bmc Genomics 14, 553 (2013).

77. Guo, Y. A. et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nature Communications 9, 1520 (2018).

78. Katainen, R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. Nature Genetics 47, 818–821 (2015).

79. Umer, H. M. et al. A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers. Human Mutation 37, 904–913 (2016).

80. Wang, K. et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet 46, 573–582 (2014).

81. Nakahashi, H. et al. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. Cell Reports 3, (2013).

82. Maurano, M. T. et al. Role of DNA Methylation in Modulating Transcription Factor Occupancy. Cell reports 12, 1184–95 (2015).

83. Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 22, 1680–8 (2012).

84. Lhoumaud, P. et al. EpiMethylTag: simultaneous detection of ATAC-seq or ChIP-seq signals with DNA methylation. Genome Biol 20, 248 (2019).

85. Schuijers, J. et al. Transcriptional Dysregulation of MYC Reveals Common Enhancer-Docking Mechanism. Cell Reports 23, 349–360 (2018).

86. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490–495 (2011).

87. Canzio, D. et al. Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin α Promoter Choice. Cell 177, 639-653.e15 (2019).

88. Schuijers, J. et al. Transcriptional Dysregulation of MYC Reveals Common Enhancer-Docking Mechanism. Cell reports 23, 349–360 (2018).

89. Barutcu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. Nature Communications 9, 1444 (2018).

90. Wang, F. et al. Long noncoding RNA HOTTIP cooperates with CCCTC-binding factor to coordinate HOXA gene expression. Biochem Bioph Res Co 500, 852–859 (2018).

91. Sun, S. et al. Jpx RNA Activates Xist by Evicting CTCF. Cell 153, 1537–1551 (2013).

92. Saldaña-Meyer, R. et al. CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. Genes & Development 28, 723–734 (2014).

93. Nora, E. P. et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930-944.e22 (2017).

94. Kubo, N. et al. Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. Biorxiv 118737 (2017) doi:10.1101/118737.

95. Rao, S. S. P. et al. Cohesin Loss Eliminates All Loop Domains. Cell 171, 305-320.e24 (2017).

96. Despang, A. et al. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nature Genetics 51, 1263–1271 (2019).

97. Ren, G. et al. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. Molecular Cell 67, 1049-1058.e6 (2017).

98. Sigova, A. A. et al. Transcription factor trapping by RNA in gene regulatory elements. Science 350, 978–981 (2015).

99. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nature Reviews Genetics 17, nrg.2016.4 (2016).

100. Gong, Y. et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nature Communications 9, 542 (2018).

101. Kuipers, E. J. et al. Colorectal cancer. Nature Reviews Disease Primers 1, nrdp201565 (2015).

102. Arnold, M. et al. Global patterns and trends in colorectal cancer incidence and mortality. Gut 66, 683 (2017).

103. Brand, M., Gaylard, P. & Ramos, J. Colorectal cancer in South Africa: An assessment of disease presentation, treatment pathways and 5-year survival. South African Medical Journal 108, 118–122 (2018).

104. Lancet, T. GLOBOCAN 2018: counting the toll of cancer. Lancet (London, England) 392, 985 (2018).

105. Rustgi, A. K. The genetics of hereditary colon cancer. Gene Dev 21, 2525–2538 (2007).

106. Mamazza, J. & Gordon, P. H. The changing distribution of large intestinal cancer. Dis Colon Rectum 25, 558–562 (1982).

107. Papagiorgis, P. Segmental distribution of some common molecular markers for colorectal cancer (CRC): influencing factors and potential implications. Tumor Biol 37, 5727–5734 (2016).

108. Roon, E. H. J. van et al. Tumour-specific methylation of PTPRG intron 1 locus in sporadic and Lynch syndrome colorectal cancer. European Journal of Human Genetics 19, 307 (2010).

109. Munkholm, P. Review article: the incidence and prevalence of colorectal cancer in inflammatory bowel disease. Alimentary Pharmacol Ther 18, 1–5 (2003).

110. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. Nat Med 21, nm.3967 (2015).

111. Poulos, R. C., Wong, Y. T., Ryan, R., Pang, H. & Wong, J. W. Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. PLOS Genetics 14, e1007779 (2018).

112. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–67 (1990).

113. Jones, S. et al. Comparative lesion sequencing provides insights into tumor evolution. Proceedings of the National Academy of Sciences of the United States of America 105, 4283–8 (2008).

114. Pino, M. S. & Chung, D. C. The Chromosomal Instability Pathway in Colon Cancer. Gastroenterology 138, 2059–2072 (2010).

115. Drost, J. et al. Sequential cancer mutations in cultured human intestinal stem cells. Nature 521, 43–47 (2015).

116. Lam, K., Pan, K., Linnekamp, J. F., Medema, J. P. & Kandimalla, R. DNA methylation based biomarkers in colorectal cancer: A systematic review. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1866, 106–120 (2016).

117. Morin, P. J. et al. Activation of beta -Catenin-Tcf Signaling in Colon Cancer by Mutations in beta -Catenin or APC. Science 275, 1787–1790 (1997).

118. Korinek, V. et al. Constitutive Transcriptional Activation by a beta -Catenin-Tcf Complex in APC-/- Colon Carcinoma. Science 275, 1784–1787 (1997).

119. Bienz, M. & Clevers, H. Linking Colorectal Cancer to Wnt Signaling. Cell 103, 311–320 (2000).

120. Orlando, F. A. et al. Aberrant crypt foci as precursors in colorectal cancer progression. J Surg Oncol 98, 207–13 (2008).

121. Downward, J. Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 3, 11–22 (2003).

122. Cox, A. D. & Der, C. J. The dark side of Ras: regulation of apoptosis. Oncogene 22, 8999–9006 (2003).

123. Sasaki, T., Hiroki, K. & Yamashita, Y. The Role of Epidermal Growth Factor Receptor in Cancer Metastasis and Microenvironment. Biomed Res Int 2013, 1–8 (2013).

124. Leslie, A., Carey, F., Pratt, N. & Steele, R. The colorectal adenoma-carcinoma sequence. Brit J Surg 89, 845–860 (2002).

125. Salem, M. E. et al. Comparative molecular analyses of left-sided colon, right-sided colon, and rectal cancers. Oncotarget 8, 86356–86368 (2017).

126. Samuels, Y. et al. High Frequency of Mutations of the PIK3CA Gene in Human Cancers. Science 304, 554–554 (2004).

127. Rosty, C. et al. PIK3CA Activating Mutation in Colorectal Carcinoma: Associations with Molecular Features and Survival. Plos One 8, e65479 (2013).

128. Bendell, J. C. et al. Phase I, Dose-Escalation Study of BKM120, an Oral Pan-Class I PI3K Inhibitor, in Patients With Advanced Solid Tumors. J Clin Oncol 30, 282–290 (2012).

129. Han, D. et al. Long noncoding RNAs: Novel players in colorectal cancer. Cancer Lett 361, 13–21 (2015).

130. Yang, Y., Junjie, P., Sanjun, C. & Ma, Y. Long non-coding RNAs in Colorectal Cancer: Progression and Future Directions. J Cancer 8, 3212–3225 (2017).

131. Xu, M., Qi, P. & Du, X. Long non-coding RNAs in colorectal cancer: implications for pathogenesis and clinical application. Modern Pathol 27, 1310–1320 (2014).

132. Ehrlich, M. DNA methylation in cancer: too much, but also too little. Oncogene 21, 5400–5413 (2002).

133. Rodriguez, J. et al. Chromosomal Instability Correlates with Genome-wide DNA Demethylation in Human Primary Colorectal Cancers. Cancer Res 66, 8462–9468 (2006).

134. Hinoue, T. et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22, 271–282 (2012).

135. Esteller, M. et al. Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. Cancer Res 60, 4366–71 (2000).

136. Tse, J. W. T., Jenkins, L. J., Chionh, F. & Mariadason, J. M. Aberrant DNA Methylation in Colorectal Cancer: What Should We Target? Trends in Cancer 3, 698–712 (2017).

137. Suzuki, H. et al. IGFBP7 is a p53-responsive gene specifically silenced in colorectal cancer with CpG island methylator phenotype. Carcinogenesis 31, 342–349 (2009).

138. IIDA, S. et al. PIK3CA mutation and methylation influences the outcome of colorectal cancer. Oncol Lett 3, 565–570 (2011).

139. Goel, A. et al. Frequent Inactivation of PTEN by Promoter Hypermethylation in Microsatellite Instability-High Sporadic Colorectal Cancers. Cancer Res 64, 3014–3021 (2004).

140. Rojas, A. et al. The aberrant methylation of TSP1 suppresses TGF-β1 activation in colorectal cancer. Int J Cancer 123, 14–21 (2008).

141. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. Cell 53, 937–947 (1988).

142. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316, 1497–1502 (2007).

143. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–19 (2011).

144. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nature Biotechnology 33, 395–401 (2015).

145. Farnham, P. J. Insights from genomic profiling of transcription factors. Nat Rev Genet 10, nrg2636 (2009).

146. Chen, Y. et al. Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 9, 609 (2012).

147. Bailey, T. et al. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. Plos Comput Biol 9, e1003326 (2013).

148. Andrews, S. FastQC: A quality control tool for high throughput sequence data. Babraham Institute (2010).

149. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinform Oxf Engl 35, 421–432 (2018).

150. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

151. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).

152. Wu, T. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881 (2010).

153. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967 (2009).

154. NovoCraft. http://www.novocraft.com (n.d.).

155. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

156. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9, R137 (2008).

157. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26, 1351–1359 (2008).

158. Heinz, S. et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 38, 576–589 (2010).

159. Mahony, S. & Pugh, F. B. Protein–DNA binding in high-resolution. Crit Rev Biochem Mol 50, 269–283 (2015).

160. Xu, S., Grullon, S., Ge, K. & Peng, W. Stem Cell Transcriptional Networks, Methods and Protocols. Methods Mol Biology Clifton N J 1150, 97–111 (2014).

161. Albert, I., Wachi, S., Jiang, C. & Pugh, B. F. GeneTrack—a genomic data processing and visualization framework. Bioinformatics 24, 1305–1306 (2008).

162. Wang, L. et al. MACE: model based analysis of ChIP-exo. Nucleic Acids Res 42, e156–e156 (2014).

163. Guo, Y. et al. Discovering homotypic binding events at high spatial resolution. Bioinformatics 26, 3028–3034 (2010).

164. Rozowsky, J. et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27, 66–75 (2009).

165. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Statistical Soc Ser B Methodol 57, 289–300 (1995).

166. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. Ann Appl Statistics 5, 1752–1779 (2011).

167. Liang, K. & Keleş, S. Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics 28, 121–122 (2012).

168. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol 11, R106 (2010).

169. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010).

170. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40, 4288–4297 (2012).

171. Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol 13, R16 (2012).

172. Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform 17, 953–966 (2016).

173. Chen, L., Wang, C., Qin, Z. S. & Wu, H. A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. Bioinformatics 31, 1889–1896 (2015).

174. Wald, A. Sequential Tests of Statistical Hypotheses. Ann Math Statistics 16, 117–186 (1945).

175. Schweikert, G., Cseke, B., Clouaire, T., Bird, A. & Sanguinetti, G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. Bmc Genomics 14, 826 (2013).

176. Gretton, A., Borgwardt, K., Rasch, M. J., Scholkopf, B. & Smola, A. J. A Kernel Method for the Two-Sample Problem. (2008).

177. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25 (2010).

178. Taslim, C. et al. Comparative study on ChIP-seq data: normalization and binding pattern characterization. Bioinformatics 25, 2334–2340 (2009).

179. Bonhoure, N. et al. Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res 24, 1157–1168 (2014).

180. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform bbw023 (2016) doi:10.1093/bib/bbw023.

181. Barski, A. & Zhao, K. Genomic location analysis by ChIP-Seq. J Cell Biochem 107, 11–18 (2009).

182. Boeva, V. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Frontiers Genetics 7, 24 (2016).

183. Grant, C. E., Bailey, T. L. & Noble, W. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011).

184. Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, I. N. S. D. The Sequence Read Archive. Nucleic Acids Res 39, D19–D21 (2011).

185. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481, 389 (2012).

186. Stark, R. & version, B.-G. package. DiffBind: differential binding analysis of ChIP-Seq peak data. (2011).

187. Haider, S. et al. A bedr way of genomic interval processing. Source Code Biology Medicine 11, 14 (2016).

188. Cook, P. R. A Model for all Genomes: The Role of Transcription Factories. J Mol Biol 395, 1–10 (2010).

189. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Research 19, 24–32 (2009).

190. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development 25, 1915–1927 (2011).

191. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515 (2010).

192. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580–585 (2013).

193. Kumegawa, K. et al. A genomic screen for long noncoding RNA genes epigenetically silenced by aberrant DNA methylation in colorectal cancer. Scientific Reports 6, 26699 (2016).

194. Huang, R.-L. et al. Methylomic Analysis Identifies Frequent DNA Methylation of Zinc Finger Protein 582 (ZNF582) in Cervical Neoplasms. Plos One 7, e41060 (2012).

195. Cheng, S.-J. et al. Hypermethylated ZNF582 and PAX1 genes in oral scrapings collected from cancer-adjacent normal oral mucosal sites are associated with aggressive progression and poor prognosis of oral cancer. Oral Oncol 75, 169–177 (2017).

196. Cheng, G. et al. A cluster of long non-coding RNAs exhibit diagnostic and prognostic values in renal cell carcinoma. Aging 11, 9597–9615 (2019).

197. Hollern, D. P. et al. E2F1 Drives Breast Cancer Metastasis by Regulating the Target Gene FGF13 and Altering Cell Migration. Sci Rep-uk 9, 10718 (2019).

198. Okada, T. et al. Upregulated expression of FGF13/FHF2 mediates resistance to platinum drugs in cervical cancer cells. Sci Rep-uk 3, 2899 (2013).

199. Song, J.-J. & Li, W. MiR-10b suppresses the growth and metastasis of colorectal cancer cell by targeting FGF13. Eur Rev Med Pharmaco 23, 576–587 (2019).

200. Tong, Y., Song, Y. & Deng, S. Combined analysis and validation for DNA methylation and gene expression profiles associated with prostate cancer. Cancer Cell Int 19, 50 (2019).

201. Bublik, D. R. et al. Regulatory module involving FGF13, miR-504, and p53 regulates ribosomal biogenesis and supports cancer cell survival. Proc National Acad Sci 114, E496–E505 (2017).

202. Hu, W. et al. Negative Regulation of Tumor Suppressor p53 by MicroRNA miR-504. Mol Cell 38, 689–699 (2010).

203. Ma, F. et al. Long non-coding RNA FGF13-AS1 inhibits glycolysis and stemness properties of breast cancer cells through FGF13-AS1/IGF2BPs/Myc feedback loop. Cancer Lett 450, 63–75 (2019).

204. Lefebvre, C. et al. Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. Plos Med 13, e1002201 (2016).

205. Bretones, G. et al. Genomic Profiles of Bone Marrow (BM) Clonal Plasma Cells (PCs) Vs Circulating Tumor Cells (CTCs) and Extramedullary (EM) Plasmacytomas in Multiple Myeloma (MM). Blood 128, 4442–4442 (2016).

206. Zhang, G. et al. Whole-exome sequencing reveals frequent mutations in chromatin remodeling genes in mammary and extramammary Paget's diseases. J Invest Dermatol 139, 789–795 (2018).

207. Litchfield, K. et al. Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. Nat Commun 6, 5973 (2015).

208. Yao, W., Huang, J. & He, H. Over-expressed LOC101927196 suppressed oxidative stress levels and neuron cell proliferation in a rat model of autism through disrupting the Wnt signaling pathway by targeting FZD3. Cell Signal 62, 109328 (2019).

209. Wong, S. et al. Clinical Significance of Frizzled Homolog 3 Protein in Colorectal Cancer Patients. Plos One 8, e79481 (2013).

210. He, W. et al. The expression of frizzled-3 receptor in colorectal cancer and colorectal adenoma. J Clin Oncol Official J Am Soc Clin Oncol 29, 444 (2011).

211. Park, J. et al. BLT2, a leukotriene B4 receptor 2, as a novel prognostic biomarker of triple-negative breast cancer. Bmb Rep 51, 373–377 (2018).

212. Venerito, M. et al. Leukotriene receptor expression in esophageal squamous cell cancer and non-transformed esophageal epithelium: a matched case control study. Bmc Gastroenterol 16, 85 (2016).

213. Park, J., Jang, J.-H. & Kim, J.-H. Mediatory role of BLT2 in the proliferation of KRAS mutant colorectal cancer cells. Biochimica Et Biophysica Acta Bba - Mol Cell Res 1866, 329–336 (2018).

214. Lee, J.-W., Kim, G.-Y. & Kim, J.-H. Androgen receptor is up-regulated by a BLT2-linked pathway to contribute to prostate cancer progression. Biochem Bioph Res Co 420, 428–433 (2012).

215. Seo, J.-M. et al. Up-regulation of BLT2 is critical for the survival of bladder cancer cells. Exp Mol Medicine 43, 129–137 (2011).

216. Kim, E.-Y. et al. BLT2 promotes the invasion and metastasis of aggressive bladder cancer cells through a reactive oxygen species-linked pathway. Free Radical Bio Med 49, 1072–1081 (2010).

217. Hennig, R. et al. BLT2 is expressed in PanINs, IPMNs, pancreatic cancer and stimulates tumour cell proliferation. Brit J Cancer 99, 1064–1073 (2008).

218. Choi, J.-A. et al. Pro-survival of estrogen receptor-negative breast cancer cells is regulated by a BLT2–reactive oxygen species-linked signaling pathway. Carcinogenesis 31, 543–551 (2010).

219. Kamalakaran, S. et al. DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. Mol Oncol 5, 77–92 (2011).

220. Yu, M. et al. Expression of CIDE proteins in clear cell renal cell carcinoma and their prognostic significance. Mol Cell Biochem 378, 145–151 (2013).

221. Cho, Y. et al. Colon cancer cell apoptosis is induced by combined exposure to the n-3 fatty acid docosahexaenoic acid and butyrate through promoter methylation. Exp Biol Med 239, 302–310 (2013).

222. Fialkova, V. et al. DNA methylation as mechanism of apoptotic resistance development in endometrial cancer patients. Gen Physiol Biophys 36, 521–529 (2017).

223. Zhan, Y. et al. A Five-Gene Signature Predicts Prognosis in Patients with Kidney Renal Clear Cell Carcinoma. Comput Math Method M 2015, 1–7 (2015).

224. Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol 19, 151 (2018).

225. Hu, H.-B., Jie, H.-Y. & Zheng, X.-X. Three Circulating LncRNA Predict Early Progress of Esophageal Squamous Cell Carcinoma. Cell Physiology Biochem Int J Exp Cell Physiology Biochem Pharmacol 40, 117–125 (2016).

226. Coe, E. A. et al. The MITF-SOX10 regulated long non-coding RNA DIRC3 is a melanoma tumour suppressor. Plos Genet 15, e1008501 (2019).

227. Liu, E. et al. Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. Cell Systems 8, 446-455.e8 (2019).

228. Al-Tassan, N. A. et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci Rep-uk 5, 10442 (2015).

229. Law, P. J. et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun 10, 2154 (2019).